

# A NEW METHOD FOR ESTIMATION AND MODEL SELECTION: $\rho$ -ESTIMATION

Y. BARAUD, L. BIRGÉ, AND M. SART

**ABSTRACT.** The aim of this paper is to present a new estimation procedure that can be applied in various statistical frameworks including density and regression and which leads to both robust and optimal (or nearly optimal) estimators. In density estimation, they asymptotically coincide with the celebrated maximum likelihood estimators at least when the statistical model is regular enough and contains the true density to estimate. For very general models of densities, including non-compact ones, these estimators are robust with respect to the Hellinger distance and converge at optimal rate (up to a possible logarithmic factor) in all cases we know. In the regression setting, our approach improves upon the classical least squares in many respects. In simple linear regression for example, it provides an estimation of the coefficients that are both robust to outliers and simultaneously rate-optimal (or nearly rate-optimal) for a large class of error distributions including Gaussian, Laplace, Cauchy and uniform among others.

## 1. INTRODUCTION

The primary scope of this paper was to design a new and more or less universal estimation method for the regression framework where we observe  $n$  independent real random variables  $X_1, \dots, X_n$  of the form  $X_i = f_i + \varepsilon_i$  where the  $f_i$  are the unknown parameters of interest and the  $\varepsilon_i$  i.i.d. real random errors with a partially unknown distribution which may be quite different from the usual Gaussian one. The problem arose from a question by Oleg Lepski to the first author during his visit to Nice in January 2012. This question was about the regression framework when the errors have rather unusual distributions, in which case the classical least squares method can be far from optimal. That was the starting point of our study which finally resulted in a much broader approach and the design of a new class of estimators with several remarkable and partly unexpected properties.

The regression frameworks that we shall consider here are of the form  $Z_i = f(W_i) + \varepsilon_i$  for  $1 \leq i \leq n$ , where the  $Z_i$  are real observations, the  $\varepsilon_i$  i.i.d. errors with density  $p$  with respect to the Lebesgue measure  $\mu$  on  $\mathbb{R}$ ,  $f$  is an unknown function from  $\mathcal{W}$  to  $\mathbb{R}$  and the  $W_i \in \mathcal{W}$  are explanatory variables which may either be deterministic, in which case  $W_i = x_i$  and  $f(x_i) = f_i$ , or random and i.i.d. This leads to the two classical regression frameworks on  $\mathbb{R}^n$  that we shall consider in the sequel:

$$X_i = f_i + \varepsilon_i \quad \text{and} \quad X_i = (W_i, Y_i) \quad \text{with} \quad Y_i = f(W_i) + \varepsilon_i \quad \text{for } 1 \leq i \leq n.$$

The first case corresponds to *fixed design regression* for which  $X_i$  has density  $p(\cdot - f_i)$  with respect to  $\mu$ , the second case to *random design regression* with i.i.d. random explanatory variables  $W_i$  independent of the  $\varepsilon_i$ .

Both examples can be set in the more general framework of independent observations with a distribution that may vary with  $i$  and that we shall now describe more precisely. We

---

*Date:* June 2, 2016.

observe  $n$  independent random variables  $X_1, \dots, X_n$  each  $X_i$  with an unknown distribution  $P_i$  on a measurable space  $(\mathcal{X}, \mathcal{A})$  and our aim is to use the vector  $\mathbf{X} = (X_1, \dots, X_n)$  of observations to estimate their joint distribution  $\mathbf{P} = \bigotimes_{i=1}^n P_i$ , that is to find a random approximation  $\hat{\mathbf{P}}(\mathbf{X}) = \bigotimes_{i=1}^n \hat{P}_i(\mathbf{X})$  of  $\mathbf{P}$  based on the observed variables  $X_i$ . To measure the quality of the approximation of  $\mathbf{P}$  by  $\hat{\mathbf{P}}$  we need a distance on the set of product measures on  $\mathcal{X}^n$ . It is known from Le Cam's work — see for instance Le Cam (1986) and Le Cam and Yang (1990) — that a very convenient one is that (here denoted by  $\mathbf{h}$ ) derived from the Hellinger distance  $h$  and introduced in Le Cam (1975):

$$\mathbf{h}^2 \left( \bigotimes_{i=1}^n P_i, \bigotimes_{i=1}^n Q_i \right) = \sum_{i=1}^n h^2(P_i, Q_i) = \frac{1}{2} \sum_{i=1}^n \int \left( \sqrt{dP_i} - \sqrt{dQ_i} \right)^2.$$

We recall that the Hellinger distance  $h$  is the bounded distance on the set of all probabilities on  $\mathcal{X}$  given by

$$(1) \quad h^2(R, T) = \frac{1}{2} \int \left( \sqrt{dR/d\mu} - \sqrt{dT/d\mu} \right)^2 d\mu \leq 1,$$

where  $\mu$  is an arbitrary positive measure which dominates both  $R$  and  $T$ , the result being independent of the choice of  $\mu$ . This is why one writes symbolically  $h^2(R, T) = (1/2) \int (\sqrt{dR} - \sqrt{dT})^2$ .

The distance  $\mathbf{h}$  between product measures provides an indicator of the quality of an estimator  $\hat{\mathbf{P}}$  of  $\mathbf{P}$  via their distance  $\mathbf{h}(\hat{\mathbf{P}}, \mathbf{P})$  and our aim is to design estimators  $\hat{\mathbf{P}}$  such that, with a probability close to one,  $\mathbf{h}(\hat{\mathbf{P}}, \mathbf{P})$  is as small as possible. We shall in particular often measure the quality of  $\hat{\mathbf{P}}$  by its *quadratic risk*  $\mathbb{E}_{\mathbf{P}}[\mathbf{h}^2(\hat{\mathbf{P}}(\mathbf{X}), \mathbf{P})]$  which is a bounded function of  $\mathbf{P}$  since  $\mathbf{h} \leq \sqrt{n}$ , the notation  $\mathbb{E}_{\mathbf{P}}$  meaning that  $\mathbf{X}$  has the distribution  $\mathbf{P}$ . As previously mentioned, we shall put a special emphasis on regression frameworks on  $\mathbb{R}^n$  and on the particularly simple example of a constant function  $f$ , which corresponds to a *translation family* for i.i.d. observations.

**1.1. Translation families.** The simplest case of a general regression framework  $Z_i = f(W_i) + \varepsilon_i$  occurs when the function  $f$  is constant and equal to  $\theta \in \Theta \subset \mathbb{R}$ . It also corresponds to fixed design regression with  $f_i = \theta$  for all  $i$ , in which case the observations  $X_i$  are i.i.d. with density  $p(\cdot - \theta)$  and distribution  $P_\theta$ ,  $\mathbf{P} = \mathbf{P}_\theta = P_\theta^{\otimes n}$  and  $\mathbf{h}^2(\mathbf{P}_\theta, \mathbf{P}_{\theta'}) = nh^2(P_\theta, P_{\theta'})$  for  $\theta, \theta' \in \Theta$ . When  $p$  is known, this is a parametric family with a single translation parameter  $\theta$  for which the problem is to find an estimator  $\hat{\theta}_n = \hat{\theta}_n(\mathbf{X})$  for  $\theta$  so that  $\hat{\mathbf{P}} = P_{\hat{\theta}_n}^{\otimes n}$ . For all densities  $p$  and  $\Theta$  an interval of positive length, it follows from Le Cam (1973) that, for all  $\pi \in (0, 1/2)$  and some constant  $c(\pi)$  depending on  $\pi$ ,

$$\sup_{\theta \in \Theta} \mathbf{P}_\theta \left[ h \left( P_{\hat{\theta}_n}, P_\theta \right) \geq c(\pi) n^{-1/2} \right] \geq \pi, \quad \text{whatever the estimator } \hat{\theta}_n.$$

When there exists a local relationship between the parameter distance and the corresponding Hellinger distance of the form

$$a|\theta - \theta'|^\alpha \leq h(P_\theta, P_{\theta'}) \leq A|\theta - \theta'|^\alpha \quad \text{for } |\theta - \theta'| \leq b \quad \text{with } a, A, b > 0, \quad 0 < \alpha \leq 1,$$

one cannot expect to build an estimator  $\hat{\theta}_n$  with convergence rate to the true  $\theta$  better than  $n^{-1/(2\alpha)}$  which we shall call the *optimal rate*.

In the past, various procedures have been considered for estimating  $\theta$ . Let us assume that the density  $p$  is symmetric and have a look at three among the most classical ones:

i) the *empirical mean*  $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$  which is the minimizer with respect to  $\theta$  of the squared empirical error  $\sum_{i=1}^n (X_i - \theta)^2$  (least squares estimator);

ii) the *empirical median*  $X_{(n/2)}$  or  $X_{((n+1)/2)}$  according to the parity of  $n$ , where  $X_{(i)}$  denotes the  $i$ -th element of the set  $\{X_1, \dots, X_n\}$  in ascending order;

iii) the *maximum likelihood estimator* (MLE for short) which maximizes the *likelihood function*  $\theta \mapsto \prod_{i=1}^n p(X_i - \theta)$ .

Unfortunately, none of them is really satisfactory in the sense that each one may behave quite poorly for some densities  $p$  as shown by the following examples. The empirical mean is only suitable when  $\mathbb{E}[\varepsilon_i] = 0$  and  $\mathbb{E}[\varepsilon_i^2] < +\infty$  as in the Gaussian case:  $p(x) = (2\pi\sigma^2)^{-1/2} \exp[x^2/(2\sigma^2)]$  where it reaches the optimal rate  $n^{-1/2}$ . But it fails miserably when the density  $p$  is Cauchy —  $p(x) = [\pi(1+x^2)]^{-1}$  — in which case one could use instead the empirical median and get again the optimal rate  $n^{-1/2}$ . When the density  $p = (1/2)\mathbb{1}_{[-1,1]}$  is uniform, both methods provide the rate  $n^{-1/2}$  while the MLE converges at the optimal rate  $n^{-1}$ . It also provides the rate  $n^{-1/2}$  for our two previous examples but, if  $p(x) = (1/4)|x|^{-1/2}\mathbb{1}_{[-1,1]}(x)$  the likelihood function is unbounded and the MLE does not even exist! In this case the empirical mean and median do exist but none of them provides the optimal rate which is, in this last case,  $n^{-2}$ . It follows that none of the three methods reaches the optimal rate for all possible densities  $p$ . Actually, each  $p$  requires the choice of a specific method depending on the characteristics of  $p$ .

There is, moreover, an additional problem which is due to the fact that our translation family is actually only a *model*, that is an approximation of the truth. This means that we pretend that our observations  $X_i$  are i.i.d. with density  $p(\cdot - \theta)$  and joint distribution  $P_\theta^{\otimes n}$  for some unknown parameter  $\theta \in \Theta$ , therefore dealing with the statistical model

$$(2) \quad \bar{S} = \{\mathbf{P}_\theta = P_\theta^{\otimes n}, \theta \in \Theta\} \quad \text{with} \quad (dP_\theta/d\mu) = p(\cdot - \theta),$$

although the true distribution is  $\mathbf{P} = \bigotimes_{i=1}^n P_i$ . Of course, if the distance  $\inf_{\theta \in \Theta} \mathbf{h}(\mathbf{P}, P_\theta^{\otimes n})$  from  $\mathbf{P}$  to our model is large, there is no hope to get a good estimation of  $\mathbf{P}$  by some  $P_{\hat{\theta}_n}^{\otimes n}$ . But when our model provides a reasonable approximation of  $\mathbf{P}$ , one would like to derive an estimator  $P_{\hat{\theta}_n}^{\otimes n}$  which remains close to  $\mathbf{P}$ . This is the so-called problem of *robustness* of estimators. It is known, for instance, that the replacement of the true  $p$  by an approximation  $q$ , even if  $h(p \cdot \mu, q \cdot \mu)$  is small, may considerably affect the value of the corresponding moments and makes methods based on moments estimation fail. The same phenomenon may happen with the MLE which should be used with great caution as emphasized by Le Cam (1990).

The situation does not improve when we consider more general regression problems and it is well-known that both the method of least squares (the multidimensional analogue of the empirical mean) and the MLE suffer from the same weaknesses as for translation families.

**1.2. What would be desirable?** In view of the conclusions of the previous section, a natural question arises: is it possible to build an estimator that can be simultaneously optimal (in some suitable sense) when the model is true and also *robust*, that is not too sensitive to small differences between the true distribution and the chosen model? There are actually two distinct problems to be solved simultaneously: one of optimality and one of robustness.

Let us first focus on optimality. We recall that we want to estimate an unknown distribution  $\mathbf{P}$  on  $\mathcal{X}^n$  which belongs to the set  $\mathcal{P}$  of all product distributions  $\bigotimes_{i=1}^n P_i$ , that is of all possible joint distributions for the independent random variables  $X_i$  and that we shall measure the quality of an estimator  $\hat{\mathbf{P}} = \bigotimes_{i=1}^n \hat{P}_i \in \mathcal{P}$  by its quadratic risk  $\mathbb{E}_{\mathbf{P}}[\mathbf{h}^2(\hat{\mathbf{P}}, \mathbf{P})] \leq n$ . Most of the time we shall assume some prior information on  $\mathbf{P}$ , for instance that it derives from some regression framework. We shall express this prior information by assuming that  $\mathbf{P} = \mathbf{P}_{\mathbf{s}}$  for some unknown parameter  $\mathbf{s}$  belonging to some given parameter set  $\mathcal{S}$ , often some subset of a linear space, either Euclidean (finite dimensional) or functional (infinite dimensional), with a one-to-one parametrization  $\mathbf{s} \mapsto \mathbf{P}_{\mathbf{s}}$ . It follows that the metric  $\mathbf{h}$  on  $\{\mathbf{P}_{\mathbf{s}}, \mathbf{s} \in \mathcal{S}\}$  can be transferred to  $\mathcal{S}$  and we shall write indifferently  $\mathbf{h}(\mathbf{P}_{\mathbf{t}}, \mathbf{P}_{\mathbf{u}})$  or  $\mathbf{h}(\mathbf{t}, \mathbf{u})$ . Unfortunately, in many situations, the set  $\{\mathbf{P}_{\mathbf{s}}, \mathbf{s} \in \mathcal{S}\}$  is too large for the existence of an estimator  $\hat{\mathbf{P}} = \mathbf{P}_{\hat{\mathbf{s}}}$  such that

$$\sup_{\mathbf{s} \in \mathcal{S}} \mathbb{E}_{\mathbf{s}} [\mathbf{h}^2(\mathbf{P}_{\mathbf{s}}, \mathbf{P}_{\hat{\mathbf{s}}})] = \sup_{\mathbf{s} \in \mathcal{S}} \mathbb{E}_{\mathbf{s}} [\mathbf{h}^2(\mathbf{s}, \hat{\mathbf{s}})] \quad (\text{with } \mathbb{E}_{\mathbf{s}} = \mathbb{E}_{\mathbf{P}_{\mathbf{s}}})$$

be substantially smaller than its maximal value  $n$ , so that the maximal quadratic risk does not provide a useful information on the quality of  $\hat{\mathbf{s}}$  unless one focuses on some specific values of  $\mathbf{s} \in \mathcal{S}$ .

Following Birgé (2006) and Baraud (2011) but also much earlier contributions including the sieves method of Grenander (1981) or the ones in Birgé and Massart (1997; 1998) and Barron, Birgé and Massart (1999) among many others, our approach in this paper is based on *models*, that is subsets  $\bar{\mathcal{S}}$  of  $\mathcal{S}$  of moderate size in order that there exists an estimator  $\hat{\mathbf{s}}$  such that  $\sup_{\mathbf{s} \in \bar{\mathcal{S}}} \mathbb{E}_{\mathbf{s}} [\mathbf{h}^2(\mathbf{s}, \hat{\mathbf{s}})] \ll n$ . This means that we do as if  $\mathbf{s}$  did belong to  $\bar{\mathcal{S}}$  and design estimators  $\hat{\mathbf{s}}(X_1, \dots, X_n)$  with values in  $\bar{\mathcal{S}}$ , although we do not necessarily assume that this is true. With this approach, a good indicator of the quality of the estimator  $\hat{\mathbf{s}}$  under the assumption that  $\mathbf{s}$  truly belongs to  $\bar{\mathcal{S}}$  is its maximal risk  $\sup_{\mathbf{s} \in \bar{\mathcal{S}}} \mathbb{E}_{\mathbf{s}} [\mathbf{h}^2(\mathbf{s}, \hat{\mathbf{s}})]$ , as compared to the so-called *minimax risk* over  $\bar{\mathcal{S}}$ ,  $R_M(\bar{\mathcal{S}}) = \inf_{\hat{\mathbf{s}}} \sup_{\mathbf{s} \in \bar{\mathcal{S}}} \mathbb{E}_{\mathbf{s}} [\mathbf{h}^2(\mathbf{s}, \hat{\mathbf{s}})]$  where the infimum runs over all possible estimators  $\hat{\mathbf{s}}$ . An *approximately optimal* estimation procedure  $\tilde{\mathbf{s}}$  should satisfy

$$\sup_{\mathbf{s} \in \bar{\mathcal{S}}} \mathbb{E}_{\mathbf{s}} [\mathbf{h}^2(\mathbf{s}, \tilde{\mathbf{s}})] \leq C_0 R_M(\bar{\mathcal{S}}),$$

where  $C_0$  (as well as all  $C_j$ 's with  $j \in \mathbb{N}$  that we shall introduce below) denotes a positive universal constant (independent of  $n$  and  $\bar{\mathcal{S}}$  and, ideally, not large).

Nevertheless, since there is no way to check precisely whether the true parameter value  $\mathbf{s}$  does actually belong to  $\bar{\mathcal{S}}$ , one would like that the previous bound remains approximately true if the model  $\bar{\mathcal{S}}$  is slightly misspecified, that is when  $\mathbf{s} \notin \bar{\mathcal{S}}$  but  $\mathbf{h}(\mathbf{s}, \bar{\mathcal{S}}) = \inf_{\mathbf{t} \in \bar{\mathcal{S}}} \mathbf{h}(\mathbf{s}, \mathbf{t})$  is small, in which case the estimator is robust. It is clear that whatever the estimator  $\hat{\mathbf{s}} \in \bar{\mathcal{S}}$ ,  $\mathbb{E}_{\mathbf{s}} [\mathbf{h}^2(\mathbf{s}, \hat{\mathbf{s}})] \geq \inf_{\mathbf{t} \in \bar{\mathcal{S}}} \mathbf{h}^2(\mathbf{s}, \mathbf{t})$ . In view of this fact and the definition of the minimax risk, an approximately optimal and robust estimator  $\tilde{\mathbf{s}}$  based on the model  $\bar{\mathcal{S}}$  should satisfy

$$(3) \quad \mathbb{E}_{\mathbf{s}} [\mathbf{h}^2(\mathbf{s}, \tilde{\mathbf{s}})] \leq C_1 \max \left\{ R_M(\bar{\mathcal{S}}), \inf_{\mathbf{t} \in \bar{\mathcal{S}}} \mathbf{h}^2(\mathbf{s}, \mathbf{t}) \right\} \quad \text{for all } \mathbf{s} \in \mathcal{S}.$$

As already mentioned, most popular methods of estimation, in particular those based on moments estimation or the MLE, are not robust and minimum contrast estimators based on the  $\mathbb{L}_2$ -contrast as well. As to classical methods which do possess some robustness properties with respect to misspecification, like methods based on the  $\mathbb{L}_1$ -contrast in regression or quantile estimation, they may unfortunately lead to sub-optimal rates of estimation.

**1.3. The search for robust and optimal estimators.** Even for the simple case of a translation parameter, finding robust estimators is definitely not obvious. An old result in this direction is from P. Huber (1964). An important research activity about robustness developed in the 60's and 70's resulting in a large number of publications. For a summary, we refer the interested reader to Huber (1981).

Attempts to design “optimal” procedures of estimation in various settings have been made by Le Cam (1973; 1975), Birgé (1983; 2006), Yang and Barron (1999) or Baraud (2011) and the construction that we shall present here is in the line of these previous papers. Actually, the problem of estimating  $\theta$  in the translation model as well as many other problems in density estimation can essentially be solved, modulo some weak assumptions, by using the methods developed in these papers.

Things become more delicate when we turn to the regression framework and, more generally, to estimating the distribution of independent but not necessarily i.i.d. distributions, for which the number of unknown parameters  $f_i$ ,  $1 \leq i \leq n$ , is a priori equal to the number of observations.

**1.3.1. *T-estimators on a model.*** Birgé (2006), following ideas from Le Cam (1973; 1975) and generalizing earlier constructions of Birgé (1983; 1984), derived a general procedure for building new estimators (called T-estimators) that satisfy (3) under some compactness assumptions on the model  $\bar{S}$ . The idea is first to build a finite discretization  $S_\eta$  at scale  $\eta$  of  $\bar{S}$  (with respect to  $\mathbf{h}$ ) and then an estimator  $\tilde{\mathbf{s}}$  with values in  $S_\eta$ , based on tests between balls in the metric space  $(\mathcal{S}, \mathbf{h})$ , centered at the points of  $S_\eta$ . These are actually robust tests between the points of  $S_\eta$ , as described for instance in Birgé (2013), so that the estimator inherits from these tests its robustness properties. The performance of the estimator is driven by a function from  $(0, +\infty)$  into  $[1/2, +\infty]$ , called the *metric dimension*  $\tilde{D}_{\bar{S}}$  of  $\bar{S}$ , that characterizes the number of points of  $S_\eta$  that are contained in balls of radius  $x\eta$  for  $x \geq 2$ . As shown in Birgé (2006), with a convenient choice of  $S_\eta$  the T-estimator  $\tilde{\mathbf{s}}$  satisfies an analogue of (3), namely that, for all  $\mathbf{s} \in \mathcal{S}$ ,

$$(4) \quad \mathbb{E}_{\mathbf{s}} [\mathbf{h}^2(\mathbf{s}, \tilde{\mathbf{s}})] \leq C_3 \max\{\eta^2, \inf_{\mathbf{t} \in \bar{S}} \mathbf{h}^2(\mathbf{s}, \mathbf{t})\} \quad \text{if } \eta \text{ satisfies } \eta^2 \geq C_2 \tilde{D}_{\bar{S}}(\eta).$$

This implies in particular that  $R_M(\bar{S}) \leq C_3 \eta^2$  and that (3) holds provided that  $R_M(\bar{S}) \geq C_4 \eta^2 > 0$ . In particular, if the function  $\tilde{D}_{\bar{S}}$  is bounded by the constant  $\overline{D}_{\bar{S}}$  (the *finite dimensional case*), one can set  $\eta^2 = C_2 \overline{D}_{\bar{S}}$  and the minimax risk  $R_M(\bar{S})$  is bounded by  $C_5 \overline{D}_{\bar{S}}$ .

**1.3.2. *Several models.*** As we can immediately see from (4) the choice of the model is crucial for the performance of  $\tilde{\mathbf{s}}$  at a given parameter  $\mathbf{s}$ . A good model should have a small dimension and be close to  $\mathbf{s}$ . Unfortunately, since  $\mathbf{s}$  is unknown, choosing a good model from scratch is possible only under rather precise information on  $\mathbf{s}$ . The solution provided by T-estimators is to deal with a large family  $\bar{\bar{S}}$  of models  $\bar{S}$  and extend the construction of T-estimators to the union of all models contained in  $\bar{\bar{S}}$ . This is precisely what has been done in Birgé (2006), resulting in the following risk bound in the case of models of finite dimension,

$$\mathbb{E}_{\mathbf{s}} [\mathbf{h}^2(\mathbf{s}, \tilde{\mathbf{s}})] \leq C_6 \inf_{\bar{S} \in \bar{\bar{S}}} \max\{\overline{D}_{\bar{S}}, \Delta(\bar{S}), \inf_{\mathbf{t} \in \bar{S}} \mathbf{h}^2(\mathbf{s}, \mathbf{t})\},$$

where the weight function  $\Delta$  satisfies  $\sum_{\bar{S} \in \bar{\bar{S}}} \exp[-\Delta(\bar{S})] \leq 1$ . As compared to (4) we see that we get the same risk bound as the one corresponding to the best model, apart from

the extra  $\Delta(\bar{\mathcal{S}})$  term which describes the complexity of the family  $\bar{\mathcal{S}}$ . The larger this family, the larger the weights  $\Delta(\bar{\mathcal{S}})$ . For simple families, one can choose  $\Delta(\bar{\mathcal{S}}) \leq C_7 \bar{D}_{\bar{\mathcal{S}}}$  and only loose a constant factor as compared to (4). Otherwise there is some additional loss which is sometimes unavoidable. The advantage of model selection is that it allows to handle many models simultaneously with the hope that one of them will be quite suitable for the estimation of the unknown parameter  $\mathbf{s}$ . For a detailed discussion about model selection, we refer the reader to Barron, Birgé and Massart (1999) or Birgé and Massart (2007).

**1.3.3. A history of dimensions.** The notion of metric dimension  $\tilde{D}_{\bar{\mathcal{S}}}$  actually applies to subsets of any metric space, not only to  $(\mathcal{S}, \mathbf{h})$ , and it is actually the right notion that is needed to control the performance of T-estimators. It is one possible way of measuring the massiveness of a model  $\bar{\mathcal{S}}$  but definitely not the only one. Others have been developed earlier, the simplest one being the ordinary dimension of a Euclidean space, but one can also mention Kolmogorov's entropy — see Kolmogorov and Tikhomirov (1961) — among other possible notions. The fact that there is often some close relationship between the minimax risk over  $\bar{\mathcal{S}}$  and some notion of dimension of  $\bar{\mathcal{S}}$  has been known for a long time, the simplest example being the estimation of the mean of a Gaussian vector with identity covariance matrix when this mean is assumed to belong to a  $D$ -dimensional linear space  $\bar{\mathcal{S}}$ . Similar results hold for parametric statistical estimation problems which are regular enough. Upper bounds for the minimax risk based on some earlier (more restrictive) version of metric dimension were developed by Le Cam (1973; 1975) and generalized by Birgé (1983) together with the connection to lower bounds previously developed by Ibragimov and Has'minskii (1980). The performance of the MLE on a parameter set  $\bar{\mathcal{S}}$  may also be deduced from some suitable notion of dimension, namely entropy with bracketing — see van de Geer (1995) and Birgé and Massart (1993) — and the concentration of the posterior distribution in Bayesian frameworks as well — see Ghosal, Gosh and van der Vaart (2000) —.

The superiority of the notion of metric dimension is due to the fact that it is a weaker notion than entropy. For instance, the metric dimension of a Euclidean space is roughly equal to its ordinary dimension while its entropy is infinite. The entropy of a compact set automatically controls its metric dimension while the reciprocal is not true. Nevertheless it is not possible to characterize the minimax risk over a model  $\bar{\mathcal{S}}$  by its metric dimension and we do not know of any notion  $\mathcal{D}_{\bar{\mathcal{S}}}$  such that

$$(5) \quad c\varphi(\mathcal{D}_{\bar{\mathcal{S}}}) \leq R_M(\bar{\mathcal{S}}) \leq C\varphi(\mathcal{D}_{\bar{\mathcal{S}}}) \quad \text{with } 0 < c < C,$$

for some suitable function  $\varphi$ , at least under very mild assumptions on  $\bar{\mathcal{S}}$ .

**1.4. From T- to  $\rho$ -estimators.** The construction of estimators from tests between balls centered on the points of some finite set, which is due to Le Cam (1973), has been extended to countable sets and developed at length in the form of T-estimators by Birgé (2006). Then, in an attempt to build a procedure for selecting estimators, Baraud (2011) designed a new method which amounts to replacing tests between balls centered at points  $\mathbf{t}$  and  $\mathbf{u}$  in  $\mathcal{S}$  by tests that tend to decide which of the two distances  $\mathbf{h}(\mathbf{t}, \mathbf{s})$  or  $\mathbf{h}(\mathbf{u}, \mathbf{s})$  is smaller, where  $\mathbf{s}$  denotes the true parameter. When applied to the discretized models  $S_\eta$  used for the construction of T-estimators, Baraud's estimators can be viewed as a particular version of T-estimators, but this alternative construction allows to relax some of the assumptions needed for the use of T-estimators. The procedure has been taken back later by Sart (2014; 2015) in a context of dependent data. A modification of Baraud's construction, following



an idea of Sart, finally led to the construction of  $\rho$ -estimators that we shall present here. It is an attempt not only to answer the above mentioned question of Oleg Lepski but also to solve this search for a “universal” estimator, at least in the case of independent observations.

In order to give a brief account of our new procedure, let us consider the problem of density estimation for i.i.d. observations  $X_1, \dots, X_n$  with values in a measurable space  $(\mathcal{X}, \mathcal{A})$ , in which case  $\mathbf{s} = (s, \dots, s)$  where  $s$  denotes the common density of the  $X_i$  with respect to some dominating measure  $\mu$  so that  $\mathbf{P}_{\mathbf{s}} = (s \cdot \mu)^{\otimes n}$  and, for  $\mathbf{t}, \mathbf{u} \in \mathcal{S}$ ,

$$(6) \quad \frac{1}{n} \mathbf{h}^2(\mathbf{t}, \mathbf{u}) = h^2(t, u) = \frac{1}{2} \int_{\mathcal{X}} (\sqrt{t} - \sqrt{u})^2 d\mu = 1 - \int_{\mathcal{X}} \sqrt{tu} d\mu = 1 - \rho(t, u),$$

where  $\rho(t, u)$  is called the *Hellinger affinity* between  $t$  and  $u$ . Let us start with a model  $\overline{\mathcal{S}}$  and two distinct densities  $t_0$  and  $t_1$  in  $\overline{\mathcal{S}}$  (that may be different from the true one  $s$ ). The difference  $\rho(s, t_1) - \rho(s, t_0) = h^2(s, t_0) - h^2(s, t_1)$  tells us which of the points  $t_0$  or  $t_1$  is closer to  $s$  with respect to the Hellinger distance. If we have at hand a good estimator  $T_n(t_0, t_1)$  of  $\rho(s, t_1) - \rho(s, t_0)$ , it can be used not only to decide which of  $t_0$  and  $t_1$  is closer to  $s$  but also, considering  $\sup_{t \in \overline{\mathcal{S}}} T_n(t_0, t)$  as an estimator of

$$T(t_0) = \sup_{t \in \overline{\mathcal{S}}} [h^2(s, t_0) - h^2(s, t)] = h^2(s, t_0) - \inf_{t \in \overline{\mathcal{S}}} h^2(s, t),$$

to see whether  $t_0$  is likely to be almost a closest point to  $s$  in  $\overline{\mathcal{S}}$ . Indeed, the smaller the quantity  $T(t_0)$ , the better  $t_0$  as an approximation of  $s$  in  $\overline{\mathcal{S}}$ . It seems therefore natural to try to minimize  $\sup_{t \in \overline{\mathcal{S}}} T_n(t_0, t)$  with respect to  $t_0 \in \overline{\mathcal{S}}$  in order to derive a good estimator of  $s$  within  $\overline{\mathcal{S}}$  provided that  $T_n(t_0, t_1)$  is close enough to  $\rho(s, t_1) - \rho(s, t_0)$  for all  $t_0, t_1$ . These are, roughly speaking, the ideas behind the construction of what we shall call a  $\rho$ -estimator since it is based on a suitable estimation of the Hellinger affinities between the true density and the points in the model.

While the study of T-estimators mainly relies on combinatorial arguments, the study of  $\rho$ -estimators involves empirical processes techniques for which a lot of results are known.

**1.5. What’s new here?** Our paper, although initially motivated by Oleg Lepski’s question and the will of finding a generic treatment of fixed-design regression under very weak assumptions (in particular no boundedness restrictions and no moment conditions), also results in both an improvement over T-estimators and a path in the direction of solving the problem summarized by (5).

It happens, as already shown in Birgé (1983), that in many situations, a lower bound on the minimax risk  $R_M(\overline{\mathcal{S}})$  over  $\overline{\mathcal{S}}$ , as defined in Section 1.2, of the form

$$R_M(\overline{\mathcal{S}}) \geq C_4 \eta^2 \quad \text{for some } \eta^2 \geq C_8 \tilde{D}_{\overline{\mathcal{S}}}(\eta)$$

actually holds, providing a reciprocal to (4); unfortunately this is not always the case. There are situations for which  $\tilde{D}_{\overline{\mathcal{S}}}(\eta) = +\infty$  for all  $\eta > 0$  and this typically happens when the diameter of  $\overline{\mathcal{S}}$  is  $\sqrt{n}$ , in particular when  $\overline{\mathcal{S}}$  is the translation model described by (2) with parameter space  $\Theta = \mathbb{R}$ . The use of a T-estimator therefore requires that  $\theta$  belong to some known interval  $[a, a + M]$  and its risk bound would unfortunately deteriorate as  $M$  becomes larger. This difficulty can be fixed via the use of a preliminary estimator, a quantile estimator for instance, allowing to locate the parameter  $\theta$  approximately but this solution does not extend to the regression framework. There are also cases with i.i.d. random variables where the quantity  $\eta^{-2} R_M(\overline{\mathcal{S}})$  with  $\eta^2 = C_2 \tilde{D}_{\overline{\mathcal{S}}}(\eta)$  tends to zero when

the number of observations tends to infinity which means that the risk bound (4) derived from the metric dimension has not the right order of magnitude. We shall even exhibit in Section 6.5 an example of a statistical model for which the metric dimension is infinite, hence the construction of a T-estimator is impossible, while a  $\rho$ -estimator reaches the optimal rate of convergence, namely  $1/\sqrt{n}$ , with respect to the Hellinger loss.

While  $\rho$ -estimators retain all the nice properties of T-estimators, in particular their robustness, their risk is bounded via new notions of dimensions which improve the one of metric dimension as shown by Corollary 16 below. These dimensions can actually be suitably controlled for many non-compact models which is an essential property for the statistical problems we want to solve.

An additional attractive feature of  $\rho$ -estimators in density estimation lies in the fact that when  $n$  is large enough, they recover the usual MLE at least when the model is parametric, regular enough and contains the true density to estimate. Some simulations developed by Sart show that this occurs even for moderate values of  $n$ . Another connection with the MLE lies in the fact that the risk bounds obtained for the MLE under bracketing entropy assumptions are still valid (up to possible numerical constants) for bounding the risks of  $\rho$ -estimators.

In the regression framework, our procedure improves upon the classical least squares from numerous aspects. First of all, we can deal with errors bearing no finite moments of any order such as the Cauchy distribution while the least squares approach cannot. Besides, we can handle various types of errors possibly leading to faster rates of estimation of the parameters than the ones reached by the least squares. Even in the case of the simple linear regression, our method may estimate at a much faster rate than the least squares, when the errors are uniformly distributed on  $[-1, 1]$  for instance. Finally, our procedure guarantees robustness properties for the resulting estimator that the use of least squares does not.

Our procedure also substantially improves upon T-estimation. A first drawback of T-estimation lies in the fact that it requires that the supremum norm of the regression function be known. When the design is random, T-estimation also requires that its distribution be known in order to achieve the properties of robustness and optimality described above. These two assumptions are unfortunately rather restrictive. In contrast, although initially conceived to handle complicated situations of regression with fixed design, our procedure also allows to deal with various random design problems and therefore handles the whole regression framework in much greater generality. More precisely,  $\rho$ -estimation does not require any knowledge about a possible bound on the regression function and about the distribution of the design, at least when the errors are modelled as symmetric. These two properties illustrate the superiority of  $\rho$ -estimation over T-estimation and we are not aware of any statistical procedure that leads to a rate-optimal estimator (up to a possible logarithmic factor) when the distribution of the errors is only assumed to belong to a large family of possible ones, including the Gaussian, Cauchy, uniform, etc.

**1.6. Connection with statistical learning theory.** The core of the proof of our main theorem relies on the control of the supremum of an empirical process, indexed by a bounded class of functions, over some vicinity of a specific element of this class. The same type of control is also needed to deal with empirical risk minimization, as explained in great details in the very nice paper by Koltchinskii (2006), and similar tools are used to handle both problems. Talagrand's concentration inequality allows to reduce the control of the supremum of the empirical process to that of its expectation and universal entropy is then



used to bound this expectation. Our use of VC-classes for bounding the universal entropy is somewhat analogous to that of Koltchinskii (2006). As a natural consequence of this parallelism, the notion of dimension that we introduce to control the risk of  $\rho$ -estimators is quite similar to the notion of local Rademacher complexity used in Koltchinskii (2006) to control the performance of empirical risk minimization. This problem and, more specifically, that of binary classification was also considered and treated with similar tools (Talagrand's theorem, universal entropy and VC-classes) in Massart and Nédélec (2006).

**1.7. Organization of the paper.** We present in Section 2 three statistical settings to which our procedure can be applied and the basic ideas underlying our approach in Section 3. The construction of the estimator and the main results about its performance on a single model can be found in Section 4. In Section 5 we show that, in favourable cases, the MLE is a particular case of  $\rho$ -estimator. We also show that the assumptions which are used to analyze the performance of the MLE in favourable situations can also be used to derive similar risk bounds for  $\rho$ -estimators. In Section 6, we illustrate the performance of  $\rho$ -estimators in the regression setting (with either fixed or random design) and provide an example for which their risk remains under control in a situation where the metric dimension of the model can be made arbitrary large and even infinite. In Section 7, we consider the problem of model selection. We establish there an oracle-type inequality and provide an application in view of estimating a regression function when the distribution of the errors belongs to a large class of densities including Laplace, Gaussian and uniform among others. We provide an annex on VC-subgraph classes in Section 8 since models  $\bar{S}$  of these types play a special role in our results. Finally, Section 9 is devoted to the proofs.

## 2. THE STATISTICAL SETTING AND EXAMPLES

**2.1. Main notations and conventions.** In the sequel we shall use the following notations and conventions. We set  $\log_+ x = \max\{\log x, 0\}$  for  $x > 0$  and  $\log_+ 0 = 0$ . For  $x, y \in \mathbb{R}$ ,  $x \wedge y$  and  $x \vee y$  denote  $\min\{x, y\}$  and  $\max\{x, y\}$  respectively,  $\delta_x$  denotes the Dirac measure at point  $x$  and  $|A|$  the cardinality of the set  $A$ . Except if otherwise specified (in Section 3 below), we shall use the conventions  $\sup \emptyset = 0$ ,  $0/0 = 1$ ,  $0 \times (+\infty) = 0$  and  $x/0 = +\infty$  for all  $x > 0$ . The word *countable* always means *finite or countable*. Throughout the paper,  $C, C', \dots$  denote positive *numerical positive* constants that may vary from line to line. The notations  $C(\cdot), C'(\cdot), \dots$  mean that  $C, C', \dots$  are positive functions depending on the argument specified in the parenthesis (when the number of arguments is too large, the dependency is specified in the text). We shall also often use the fact that

$$(7) \quad (x + y)^2 \leq (1 + \alpha)x^2 + (1 + \alpha^{-1})y^2 = (1 + \alpha)(x^2 + \alpha^{-1}y^2) \quad \text{for all } \alpha > 0.$$

Our definitions and results will actually involve a number of numerical constants. In order to avoid complicated formulas, we shall give specific names to the numerical constants that will be systematically used in the sequel.

$$(8) \quad \begin{cases} c_0 = \frac{1}{8} \left(1 - \frac{1}{\sqrt{2}}\right) = \frac{1}{8(2 + \sqrt{2})}; & c_1 = 2(7 + 4\sqrt{2}); & c'_1 = 2(c_1 - 1); \\ c_2 = 1 + \frac{1}{\sqrt{2}} = \frac{\sqrt{2} + 1}{\sqrt{2}} = \frac{1}{16c_0}; & \kappa = 357; & c_3 = 8\kappa c_2; & c_4 = 2.5c_3. \end{cases}$$

**2.2. The general statistical setting.** Let  $\mathbf{X} = (X_1, \dots, X_n)$  be a vector of independent random variables with values in a product of measured spaces  $(\prod_{i=1}^n \mathcal{X}_i, \otimes_{i=1}^n \mathcal{A}_i, \otimes_{i=1}^n \mu_i)$ . We assume that for each  $i$ ,  $X_i$  admits a density  $s_i$  with respect to  $\mu_i$  and our aim is to estimate  $\mathbf{s} = (s_1, \dots, s_n)$  from the observation of  $\mathbf{X}$ . To avoid trivialities, we shall always assume that  $n \geq 3$ . We shall emphasize the dependence of the distribution of  $\mathbf{X}$  with respect to the unknown parameter  $\mathbf{s}$  by writing  $\mathbb{P}_{\mathbf{s}}[\mathbf{X} \in A]$  for a measurable set  $A$  and  $\mathbb{E}_{\mathbf{s}}[g(\mathbf{X})]$  for an integrable function  $g$ .

On the measured space  $(\mathcal{X}_i, \mathcal{A}_i, \mu_i)$  we consider the set  $\mathcal{L}'_i$  of all measurable real-valued functions  $u$  such that  $\int_{\mathcal{X}_i} |u| d\mu_i < +\infty$  and the subset  $\mathcal{L}_i$  of  $\mathcal{L}'_i$  of all probability densities with respect to  $\mu_i$ , that is non-negative measurable functions  $u$  on  $(\mathcal{X}_i, \mathcal{A}_i)$  such that  $\int_{\mathcal{X}_i} u d\mu_i = 1$ . We equip  $\mathcal{L}_i$  with the Hellinger pseudometric  $h$  given, according to (1), by

$$h^2(u, u') = \frac{1}{2} \int_{\mathcal{X}_i} (\sqrt{u} - \sqrt{u'})^2 d\mu_i \quad \text{for all } u, u' \in \mathcal{L}_i.$$

Note that  $h$  is only a pseudometric (symmetric and satisfying the triangular inequality) since  $h(u, u') = 0$  if  $u \neq u'$  but  $u = u'$   $\mu$ -a.e., although  $h$  is a genuine distance on the corresponding probability space since  $h(u \cdot \mu, u' \cdot \mu)$  implies that  $u \cdot \mu = u' \cdot \mu$ . In particular,  $s_i$  may be any element of  $\mathcal{L}_i$  such that the distribution of  $X_i$  can be written  $s_i \cdot \mu_i$ .

We define  $\mathcal{L}_0$  as the product space  $\prod_{i=1}^n \mathcal{L}_i$ , call the elements of  $\mathcal{L}_0$  densities and equip it with the pseudometric  $\mathbf{h}$  given, by analogy with (6) and following Le Cam (1975), by

$$\mathbf{h}^2(\mathbf{t}, \mathbf{t}') = \frac{1}{2} \int (\sqrt{\mathbf{t}} - \sqrt{\mathbf{t}'})^2 d\boldsymbol{\mu} = \sum_{i=1}^n h^2(t_i, t'_i) \leq n \quad \text{for } \mathbf{t}, \mathbf{t}' \in \mathcal{L}_0.$$

For simplicity we shall still call  $h$  and  $\mathbf{h}$  *distances*, although they are only pseudometrics on  $\mathcal{L}_i$ ,  $1 \leq i \leq n$  and  $\mathcal{L}_0$  respectively, call  $(\mathcal{L}_0, \mathbf{h})$  a pseudometric space and introduce the following definition.

**Definition 1.** A subset  $S$  of  $\mathcal{L}_0$  is said to be identifiable if, when  $\mathbf{u}, \mathbf{u}' \in S$  are such that  $\mathbf{u} \neq \mathbf{u}'$ , then  $\mathbf{h}(\mathbf{u}, \mathbf{u}') > 0$  or, equivalently, if  $\mathbf{h}$  is a genuine distance on  $S$ .

For  $\mathbf{u} \in \mathcal{L}_0$  and more generally for  $\mathbf{u} \in \prod_{i=1}^n \mathcal{L}'_i$ , we shall set

$$\mathbf{u}(\mathbf{X}) = \sum_{i=1}^n u_i(X_i) \quad \text{and} \quad \int \mathbf{u} d\boldsymbol{\mu} = \sum_{i=1}^n \int_{\mathcal{X}_i} u_i d\mu_i.$$

Hereafter, we shall deal with estimators with values in  $\mathcal{L}_0$  and measure their performances by the risk induced by the loss function  $\mathbf{h}^2$ . For simplicity we shall also call  $\mathbf{h}$  the Hellinger distance and define the Hellinger affinity  $\boldsymbol{\rho}$  between two elements  $\mathbf{t}$  and  $\mathbf{t}'$  of  $\mathcal{L}_0$  as

$$\boldsymbol{\rho}(\mathbf{t}, \mathbf{t}') = \int \sqrt{\mathbf{t}\mathbf{t}'} d\boldsymbol{\mu} = \sum_{i=1}^n \int_{\mathcal{X}_i} \sqrt{t_i t'_i} d\mu_i = n - \mathbf{h}^2(\mathbf{t}, \mathbf{t}') \geq 0.$$

For  $\mathbf{t} \in \mathcal{L}_0$  and  $y > 0$ , we shall denote by  $\mathcal{B}(\mathbf{t}, y)$  the closed ball of center  $\mathbf{t}$  and radius  $y$  in the pseudometric space  $(\mathcal{L}_0, \mathbf{h})$  and, given some subset  $S$  of  $\mathcal{L}_0$ , by  $\mathcal{B}^S(\mathbf{s}, y) = \mathcal{B}(\mathbf{s}, y) \cap S$  the closed Hellinger ball in  $S$  centered at  $\mathbf{s}$  with radius  $y$ . Finally,  $\mathbf{h}(\mathbf{t}, S) = \inf_{\mathbf{t}' \in S} \mathbf{h}(\mathbf{t}, \mathbf{t}')$ .

This general setting will allow us to deal in particular with the three following specific frameworks.

**2.3. The density framework.** In this framework, we assume that the random variables  $X_1, \dots, X_n$  are i.i.d. with values in a measured space  $(\mathcal{X}, \mathcal{A}, \mu)$  and common density  $s$  with respect to  $\mu$ , which leads to  $\mathcal{X}_i = \mathcal{X}$ ,  $\mu_i = \mu$  for all  $i$  and  $\mathbf{s} = (s, \dots, s)$ . In this particular context, it will be convenient to identify a density  $t$  on  $(\mathcal{X}, \mathcal{A}, \mu)$  with the element  $\mathbf{t} = (t, t, \dots, t)$  of  $\mathcal{L}_0$ , which we shall do in the sequel. Given two densities  $t, t'$  on  $(\mathcal{X}, \mathcal{A}, \mu)$ , we have the relations

$$\mathbf{h}^2(\mathbf{t}, \mathbf{t}') = nh^2(t, t') \quad \text{and} \quad \boldsymbol{\rho}(\mathbf{t}, \mathbf{t}') = n\rho(t, t').$$

The risk of an estimator  $\tilde{\mathbf{s}} = (\tilde{s}, \dots, \tilde{s})$  of  $\mathbf{s}$  is therefore  $\mathbb{E}_{\mathbf{s}} [\mathbf{h}^2(\mathbf{s}, \tilde{\mathbf{s}})] = n\mathbb{E}_{\mathbf{s}} [h^2(s, \tilde{s})]$ .

**2.4. The homoscedastic regression framework with fixed design.** In this framework, we assume that the  $X_i$  are real-valued random variables satisfying equations of the form

$$X_i = f_i + \lambda \varepsilon_i \quad \text{for } i = 1, \dots, n, \quad \lambda > 0,$$

where the vector  $\mathbf{f} = (f_1, \dots, f_n)$  belongs to  $\mathbb{R}^n$ , the  $\varepsilon_i$  are real-valued i.i.d. random variables with density  $p$  with respect to the *Lebesgue measure*  $\mu$  on  $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ ,  $\mathcal{B}(A)$  denoting the Borel  $\sigma$ -algebra on the topological space  $A$ . It follows that the density of  $X_i$  is  $s_i(x) = \lambda^{-1}p(\lambda^{-1}(x - f_i))$  so that estimating  $s_i$  amounts to estimating  $\lambda$ ,  $p$  and  $f_i$ . Our aim is therefore to estimate  $\mathbf{f}$ ,  $p$  and  $\lambda$  from the observation of  $X_1, \dots, X_n$ . To deal with this framework, it will be convenient to introduce the following notations: for  $f$  and  $x$  in  $\mathbb{R}$ ,  $\lambda$  in  $\mathbb{R}_+ \setminus \{0\}$ ,  $\mathbf{f}$  and  $\mathbf{x} = (x_1, \dots, x_n)$  in  $\mathbb{R}^n$ , we set

$$(9) \quad p_{f,\lambda}(x) = \frac{1}{\lambda}p\left(\frac{x-f}{\lambda}\right) \quad \text{for all } x \in \mathbb{R}; \quad p_f = p_{f,1};$$

$$(10) \quad \mathbf{p}_{\mathbf{f},\lambda}(\mathbf{x}) = \left(\frac{1}{\lambda}p\left(\frac{x_1-f_1}{\lambda}\right), \dots, \frac{1}{\lambda}p\left(\frac{x_n-f_n}{\lambda}\right)\right) \quad \text{for all } \mathbf{x} \in \mathbb{R}^n \quad \text{and} \quad \mathbf{p}_{\mathbf{f}} = \mathbf{p}_{\mathbf{f},1}.$$

It follows that  $p_f(x) = p(x - f)$ ,  $p_{0,\lambda}(x) = \lambda^{-1}p(x/\lambda)$ , etc. In this framework we take, for  $i = 1, \dots, n$ ,  $\mathcal{X}_i = \mathbb{R}$ ,  $\mathcal{A}_i = \mathcal{B}(\mathbb{R})$ ,  $\mu_i = \mu$  and  $s_i = p_{f_i,\lambda}$  so that the density of  $\mathbf{X}$  is  $\bigotimes_{i=1}^n p_{f_i,\lambda}$  and therefore entirely determined by  $\mathbf{p}_{\mathbf{f},\lambda}$ .

**2.5. The homoscedastic regression framework with random design.** Let  $(W, Y)$  be a pair of random variables with values in  $(\mathcal{W} \times \mathbb{R}, \mathcal{W} \otimes \mathcal{B}(\mathbb{R}))$  linked by the relation

$$(11) \quad Y = f(W) + \varepsilon$$

where  $f$  is unknown in a set  $\mathcal{F}$  of measurable functions from  $(\mathcal{W}, \mathcal{W})$  into  $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$  and  $\varepsilon$  is an unobservable random variable, independent of  $W$  and admitting a *known* (or *approximately known*) density  $p$  with respect to the *Lebesgue measure*  $\mu$ . In contrast, the distribution  $\nu$  of  $W$  is possibly unknown.

Our aim is to estimate  $f$ , or equivalently the conditional distribution of  $Y$  given  $W$ , from the observation of  $\mathbf{X} = (X_1, \dots, X_n)$  where the  $X_i$  are i.i.d. with the same distribution on  $(\mathcal{W} \times \mathbb{R})$  as that of the pair  $(W, Y)$  so that  $\mathbf{X}$  can be identified to  $(\mathbf{W}, \mathbf{Y})$  with  $\mathbf{W} = (W_1, \dots, W_n)$  and  $\mathbf{Y} = (Y_1, \dots, Y_n)$ . Since the density of  $(W, Y)$  with respect to the dominating measure  $\nu \otimes \mu$  is  $s(w, y) = p(y - f(w)) = p_f(w, y)$ ,  $p_f(W, \cdot)$  is the conditional density of  $Y$  given  $W$  with respect to  $\mu$ . It is therefore natural to look for estimators of  $s$  of the form  $\hat{s} = p_{\hat{f}}$  where  $\hat{f} = \hat{f}(\mathbf{W}, \mathbf{Y})$  is an estimator of  $f$  which also provides an estimator  $p_{\hat{f}}(w, \cdot)$  of the conditional density of  $\mathbf{Y}$  when  $\mathbf{W} = w$ . At this stage, it is important to emphasize the fact that the construction of  $\hat{s}$  *should not involve*  $\nu$  in order to provide

genuine estimators of  $f$  and  $p_f$ . As we shall see in Section 6.3, the  $\rho$ -estimator of  $s$  derived from our general method does satisfy this requirement.

To evaluate the performance of  $\hat{f}(\mathbf{W}, \mathbf{Y})$ , we use the risk  $\mathbb{E}_{\mathbf{s}} [\mathbf{h}^2(\mathbf{s}, \hat{\mathbf{s}})]$  of  $\hat{\mathbf{s}}(\mathbf{W}, \mathbf{Y})$  or, equivalently, the risk of  $\mathbf{p}_{\hat{\mathbf{f}}}$  which writes

$$\mathbb{E}_{\mathbf{s}} [\mathbf{h}^2(\mathbf{p}_{\mathbf{f}}, \mathbf{p}_{\hat{\mathbf{f}}})] = n \mathbb{E}_{\mathbf{s}} [h^2(p_f, p_{\hat{f}})] = n \mathbb{E}_{\mathbf{s}} \left[ \int_{\mathcal{W}} h^2(p_f(w, \cdot), p_{\hat{f}}(w, \cdot)) d\nu(w) \right].$$

### 3. BASIC IDEAS UNDERLYING OUR APPROACH

**3.1. The density framework.** The aim of this section is to present the basic ideas and formulas underlying our approach. For the sake of simplicity, we shall restrict this introduction to the density framework described in Section 2.3 where the observations  $X_1, \dots, X_n$  are i.i.d. with an unknown density  $s$  with respect to  $\mu$ .

Given two candidate densities  $t, t'$  for  $s$ , one should prefer  $t'$  to  $t$  if it is closer to  $s$ , that is, if  $h^2(s, t')$  is smaller than  $h^2(s, t)$  or equivalently if  $\rho(s, t') - \rho(s, t) > 0$ . Deciding whether  $t'$  is preferable to  $t$  amounts thus to estimating the difference  $\rho(s, t') - \rho(s, t)$  in a suitable way. To do so, we start by an approximation of the affinity  $\rho$ . For two densities  $t$  and  $t'$ , we set

$$(12) \quad r = \frac{t + t'}{2} \quad \text{and} \quad \varrho(s, t, t') = \frac{1}{2} \left[ \rho(t, r) + \int_{\mathcal{X}} \sqrt{\frac{t}{r}} s d\mu \right] < +\infty,$$

using the special convention that  $t/r = 0$  when  $t = t' = r = 0$ . It was proved in Proposition 1 of Baraud (2011) that

$$(13) \quad 0 \leq \varrho(s, t, t') - \rho(s, t) \leq [h^2(s, t) + h^2(s, t')] / \sqrt{2}.$$

The important point about (13) lies in the fact that the constant  $\sqrt{2}$  is larger than 1. This makes it possible to use the sign of the difference

$$T(s, t, t') = \varrho(s, t', t) - \varrho(s, t, t')$$

as an alternative benchmark to find which of  $t$  and  $t'$  is closer to  $s$  (up to a multiplicative constant). It actually follows from (13), as shown in Corollary 1 in Baraud (2011), that

$$(14) \quad T(s, t, t') \leq \left(1 + \frac{1}{\sqrt{2}}\right) h^2(s, t) - \left(1 - \frac{1}{\sqrt{2}}\right) h^2(s, t') = c_2 h^2(s, t) - 8c_0 h^2(s, t')$$

and

$$(15) \quad T(s, t, t') \geq \left(1 - \frac{1}{\sqrt{2}}\right) h^2(s, t) - \left(1 + \frac{1}{\sqrt{2}}\right) h^2(s, t') = 8c_0 h^2(s, t) - c_2 h^2(s, t').$$

Given some subset  $S$  of  $\mathcal{L}_0$ , (14) and the fact that  $T(s, t, t) = 0$  also imply that, for  $t \in S$

$$0 \leq \sup_{t' \in S} T(s, t, t') \leq c_2 h^2(s, t) - 8c_0 h^2(s, S) \quad \text{and} \quad 0 \leq \inf_{t \in S} \sup_{t' \in S} T(s, t, t') \leq \sqrt{2} h^2(s, S).$$

If  $u \in S$  is such that

$$h^2(s, u) > (8c_0)^{-1} (c_2 + \sqrt{2}) h^2(s, S) = (5 + 4\sqrt{2}) h^2(s, S),$$

it follows from (15) that

$$\sup_{t' \in S} T(s, u, t') \geq 8c_0 h^2(s, u) - c_2 h^2(s, S) > [(c_2 + \sqrt{2}) - c_2] h^2(s, S) = \sqrt{2} h^2(s, S),$$

and  $u$  cannot be a minimizer of  $t \mapsto \sup_{t' \in S} T(s, t, t')$ . Hence any minimizer  $\bar{s}$  of this function does satisfy  $h^2(s, \bar{s}) \leq (5 + 4\sqrt{2}) h^2(s, S) < 11h^2(s, S)$ . Therefore, minimizing over  $S$  the function  $t \mapsto \sup_{t' \in S} T(s, t, t')$  leads to some point  $\bar{s} \in S$  which, up to a factor smaller than 11, is the closest to  $s$ , that is, the best approximation of  $s$  in  $S$ . In particular, if  $s \in S$ ,  $\bar{s} = s$ .

Unfortunately,  $T(s, t, t')$  depends on  $\varrho(s, t, t')$  which depends on the unknown  $s$ . Our interest for the quantity  $\varrho(s, t, t')$  rather than  $\rho(s, t)$  lies in the fact that the former can be estimated by its empirical counterpart, namely

$$(16) \quad \varrho(\mathbf{X}, t, t') = \frac{1}{2n} \sum_{i=1}^n \left[ \rho(t, r) + \sqrt{\frac{t}{r}}(X_i) \right] \quad \text{with } r = \frac{t + t'}{2},$$

which is an unbiased estimator of  $\varrho(s, t, t')$ . A natural way of deciding which of the densities  $t$  or  $t'$  is the closest to  $s$  is therefore to replace the unknown  $T(s, t, t')$  by an unbiased estimator, namely the statistic

$$T(\mathbf{X}, t, t') = \varrho(\mathbf{X}, t', t) - \varrho(\mathbf{X}, t, t').$$

Note that

$$(17) \quad T(\mathbf{X}, t, t') = \frac{1}{2} [\rho(t', r) - \rho(t, r)] + \frac{1}{\sqrt{2}n} \sum_{i=1}^n \psi \left( \sqrt{\frac{t'}{t}}(X_i) \right),$$

where  $\psi$  is the Lipschitz, increasing function from  $[0, +\infty]$  to  $[-1, 1]$  (with Lipschitz constant not larger than 1.143) given by

$$(18) \quad \psi(u) = \sqrt{\frac{1}{1+u^{-2}}} - \sqrt{\frac{1}{1+u^2}} = \frac{u-1}{\sqrt{1+u^2}} \quad \text{for } u \in [0, +\infty) \quad \text{and} \quad \psi(+\infty) = 1.$$

Here we use the convention that  $t'(X_i)/t(X_i) = 1$  when  $t(X_i) = t'(X_i) = 0$  as indicated in Section 2.1. This convention is indeed consistent with the one we started from on the ratio  $t/r$  since when  $t(X_i) = t'(X_i) = r(X_i) = 0$  for some  $i$ ,

$$\left[ \sqrt{\frac{t'}{r}}(X_i) - \sqrt{\frac{t}{r}}(X_i) \right] = 0 - 0 = 0 \quad \text{and} \quad \psi \left( \sqrt{\frac{t'}{t}}(X_i) \right) = \psi \left( \frac{0}{0} \right) = \psi(1) = 0.$$

Replacing the “ideal” statistic  $T(s, t, t')$  by its empirical counterpart  $T(\mathbf{X}, t, t')$  leads to an estimation error given by the process  $Z(\mathbf{X}, \cdot, \cdot)$  defined on  $\mathcal{L}_0^2$  by

$$(19) \quad \begin{aligned} Z(\mathbf{X}, t, t') &= \sqrt{2} [T(\mathbf{X}, t, t') - T(s, t, t')] \\ &= \sqrt{2} \left( [\varrho(\mathbf{X}, t', t) - \varrho(s, t', t)] - [\varrho(\mathbf{X}, t, t') - \varrho(s, t, t')] \right) \\ &= \frac{1}{n} \sum_{i=1}^n \left[ \psi \left( \sqrt{\frac{t'}{t}}(X_i) \right) - \mathbb{E}_s \left[ \psi \left( \sqrt{\frac{t'}{t}}(X_i) \right) \right] \right]. \end{aligned}$$

**3.2. The general framework.** We may similarly apply the previous reasoning to the more general context of independent but not necessarily i.i.d. variables  $X_i$ ,  $1 \leq i \leq n$ . To do so, we shall extend the previous notations to elements of  $\mathcal{L}_0$  and, in view of the application to

the regression setting, we shall not renormalize the sums by  $1/n$ . This leads to the following notations to be used throughout this paper: for  $\mathbf{t}, \mathbf{t}' \in \mathcal{L}_0$  and  $\mathbf{r} = (\mathbf{t} + \mathbf{t}')/2$ , we set

$$\begin{aligned} \varrho(\mathbf{X}, \mathbf{t}, \mathbf{t}') &= \frac{1}{2} \left[ \rho(\mathbf{t}, \mathbf{r}) + \sqrt{\frac{\mathbf{t}}{\mathbf{r}}}(\mathbf{X}) \right] = \frac{1}{2} \sum_{i=1}^n \left[ \rho(t_i, r_i) + \sqrt{\frac{t_i}{r_i}}(X_i) \right]; \\ (20) \quad \mathbf{T}(\mathbf{X}, \mathbf{t}, \mathbf{t}') &= \varrho(\mathbf{X}, \mathbf{t}', \mathbf{t}) - \varrho(\mathbf{X}, \mathbf{t}, \mathbf{t}') \end{aligned}$$

$$\begin{aligned} &= \frac{1}{2} \sum_{i=1}^n [\rho(t'_i, r_i) - \rho(t_i, r_i)] + \frac{1}{\sqrt{2}} \sum_{i=1}^n \psi \left( \sqrt{\frac{t'_i}{t_i}}(X_i) \right); \\ (21) \quad \mathbf{Z}(\mathbf{X}, \mathbf{t}, \mathbf{t}') &= \psi \left( \sqrt{\frac{\mathbf{t}'}{\mathbf{t}}}(\mathbf{X}) \right) - \mathbb{E}_{\mathbf{s}} \left[ \psi \left( \sqrt{\frac{\mathbf{t}'}{\mathbf{t}}}(\mathbf{X}) \right) \right] \\ &= \sum_{i=1}^n \left[ \psi \left( \sqrt{\frac{t'_i}{t_i}}(X_i) \right) - \mathbb{E}_{s_i} \left[ \psi \left( \sqrt{\frac{t'_i}{t_i}}(X_i) \right) \right] \right]. \end{aligned}$$

With these notations,  $\varrho(\mathbf{X}, \mathbf{t}, \mathbf{t}')$  is in fact the analogue of  $n\varrho(\mathbf{X}, t, t')$  given by (16) (and actually equal to it in the density framework), and so on.

#### 4. ESTIMATION ON A MODEL

**4.1. Models.** As already mentioned, our construction of estimators will be based on “models”. A model  $\bar{S} \subset \mathcal{L}_0$  should be viewed as an approximation set for the true unknown parameter  $\mathbf{s}$ , which is used to build an estimator. It does not necessarily contain  $\mathbf{s}$ , although we shall occasionally assume so. Typical models are either the parametric models that are used in Statistics or more general subsets of  $\mathcal{L}_0$  with well-known approximation properties that are derived from Approximation Theory in order to get a control on the approximation error  $\mathbf{h}(\mathbf{s}, \bar{S})$ . For measurability reasons to be explained later, we shall adopt the following definition for a model.

**Definition 2.** A model  $\bar{S}$  is a nonempty separable subset of the pseudometric space  $(\mathcal{L}_0, \mathbf{h})$  which means that one can find a countable subset  $S$  of  $\bar{S}$  such that  $\mathbf{h}(\mathbf{t}, S) = 0$  for all  $\mathbf{t} \in \bar{S}$ .

In typical situations,  $\mathcal{L}_0$  itself is separable so that any nonempty subset of  $\mathcal{L}_0$  can be used as a model. In the density framework described in Section 2.3, we have identified a density  $t$  on  $(\mathcal{X}, \mathcal{A}, \mu)$  with the element  $\mathbf{t} = (t, t, \dots, t)$  of  $\mathcal{L}_0$ . Similarly, we shall identify a separable set  $\bar{S}$  of densities on  $\mathcal{X}$  to the subset  $\{\mathbf{t} = (t, \dots, t), t \in \bar{S}\} \subset \mathcal{L}_0$  and for simplicity denote both sets the same way.

**4.2. Construction of the estimators.** In order to avoid measurability issues, the construction of our estimator will be performed over countable subsets  $S$  of  $\mathcal{L}_0$  only. Besides, this corresponds to the practical point of view since numerical optimization will always be done over a finite set. Replacing the original model  $\bar{S}$  by a countable and dense subset  $S$  does not increase the approximation error since then  $\mathbf{h}(\mathbf{s}, S) = \mathbf{h}(\mathbf{s}, \bar{S})$ . If instead we replace  $\bar{S}$  by  $S \subset \mathcal{L}_0$  which only satisfies  $\sup_{\mathbf{u} \in \bar{S}} \mathbf{h}(\mathbf{u}, S) = \sup_{\mathbf{u} \in \bar{S}} \inf_{\mathbf{t} \in S} \mathbf{h}(\mathbf{u}, \mathbf{t}) \leq \eta$ , this replacement may involve an additional error  $|\mathbf{h}(\mathbf{s}, S) - \mathbf{h}(\mathbf{s}, \bar{S})|$  which is not larger than  $\eta$ . We postpone the discussion about what should be suitable choices of  $S$  for a given model  $\bar{S}$  to Section 4.5.



We shall also always assume  $S$  to be *identifiable* in order that  $(S, \mathbf{h})$  be a genuine metric space. Note that this identifiability condition is not restrictive at all: if  $S'$  is countable but not identifiable, one can withdraw from it the redundant points in order to get an identifiable subset  $S \subset S'$  with  $\mathbf{h}(\mathbf{s}, S) = \mathbf{h}(\mathbf{s}, S')$  for all  $\mathbf{s}$  in  $\mathcal{L}_0$ . If  $\bar{S}$  is identifiable and  $S$  is a subset of  $\bar{S}$ , there is nothing to do. Otherwise, we proceed as indicated above. In any case, we shall always assume in the sequel and sometimes without further notice, that the countable sets  $S$  that we shall use in our construction are identifiable so that  $(S, \mathbf{h})$  is a metric space.

Given  $S$ , we noticed in Section 3.1 that, in the density framework, an almost best approximation of the density  $s$  in  $S$  can be obtained by minimizing over  $S$  the function  $t \mapsto \sup_{t' \in S} T(s, t, t')$ . If we assume that  $T(\mathbf{X}, t, t')$  provides a good approximation of  $T(s, t, t')$ , it looks natural to minimize  $\sup_{t' \in S} T(\mathbf{X}, t, t')$  with respect to  $t \in S$  in order to derive a good estimation  $\hat{s}(\mathbf{X})$  of  $s$ . This suggests the following construction in the general situation of independent random variables. For each  $\mathbf{t} \in S$  we define

$$\Upsilon(S, \mathbf{t}) = \sup_{\mathbf{t}' \in S} \mathbf{T}(\mathbf{X}, \mathbf{t}, \mathbf{t}'),$$

which is always non-negative since  $\mathbf{T}(\mathbf{X}, \mathbf{t}, \mathbf{t}) = 0$ , and

$$(22) \quad \mathcal{E}(\mathbf{X}, S) = \left\{ \tilde{\mathbf{s}} \in S \mid \Upsilon(S, \tilde{\mathbf{s}}) \leq \inf_{\mathbf{t} \in S} \Upsilon(S, \mathbf{t}) + \frac{\kappa}{10} \right\} \text{ with } \kappa \text{ given by (8).}$$

Finally, we define our estimator of  $\mathbf{s}$  as any element (chosen in a measurable way)  $\hat{\mathbf{s}}$  in the closure  $\text{Cl}(\mathcal{E}(\mathbf{X}, S))$  of  $\mathcal{E}(\mathbf{X}, S)$  in  $\mathcal{L}_0$  with respect to  $\mathbf{h}$ , that is the set

$$\{\mathbf{t} \in \mathcal{L}_0 \text{ such that } \mathbf{h}(\mathbf{t}, \mathcal{E}(\mathbf{X}, S)) = 0\}.$$

We shall call such an estimator a  $\rho$ -estimator, the greek letter  $\rho$  referring to the Hellinger affinity. Although it depends on our choice of  $S$  we shall, for simplicity, omit to make this dependence of  $\hat{\mathbf{s}}$  with respect to  $S$  explicit in our notations. Though the risk bounds we shall establish remain valid for any choice of  $\hat{\mathbf{s}}$  in  $\text{Cl}(\mathcal{E}(\mathbf{X}, S))$ , we recommend in practice to choose  $\hat{\mathbf{s}}$  as a minimizer of  $\Upsilon(S, \cdot)$  over  $S$  whenever it exists.

**4.3. Main theorem.** The properties of our estimator follow from those of the empirical process  $\mathbf{Z}(\mathbf{X}, \cdot, \cdot)$  defined by (21). Given an element  $\bar{\mathbf{s}} \in \mathcal{L}_0$  and a positive number  $y$ , we set

$$(23) \quad \mathcal{B}^S(\mathbf{s}, \bar{\mathbf{s}}, y) = \{\mathbf{t} \in S \mid \mathbf{h}^2(\mathbf{s}, \mathbf{t}) + \mathbf{h}^2(\mathbf{s}, \bar{\mathbf{s}}) \leq y^2\}$$

and

$$\mathbf{w}^S(\mathbf{s}, \bar{\mathbf{s}}, y) = \mathbb{E}_{\mathbf{s}} \left[ \sup_{\mathbf{t} \in \mathcal{B}^S(\mathbf{s}, \bar{\mathbf{s}}, y)} |\mathbf{Z}(\mathbf{X}, \bar{\mathbf{s}}, \mathbf{t})| \right],$$

with  $\mathbf{w}^S(\mathbf{s}, \bar{\mathbf{s}}, y) = 0$  if  $\mathcal{B}^S(\mathbf{s}, \bar{\mathbf{s}}, y)$  is empty, according to our convention. When  $\mathbf{s}$  belongs to  $S$  and one takes  $\bar{\mathbf{s}} = \mathbf{s}$ ,

$$\mathbf{w}^S(\mathbf{s}, \mathbf{s}, y) = \mathbb{E}_{\mathbf{s}} \left[ \sup_{\mathbf{t} \in \mathcal{B}^S(\mathbf{s}, \mathbf{s}, y)} |\mathbf{Z}(\mathbf{X}, \mathbf{s}, \mathbf{t})| \right]$$

measures, in some sense, the massiveness of  $S$  in a neighborhood of  $\mathbf{s}$ . Since  $-1 \leq \psi \leq 1$ , the process  $|\mathbf{Z}(\mathbf{X}, \cdot, \cdot)|$  is bounded by  $2n$  and the non-decreasing mapping  $y \mapsto \mathbf{w}^S(\mathbf{s}, \bar{\mathbf{s}}, y)$  as well. This implies that the number

$$D^S(\mathbf{s}, \bar{\mathbf{s}}) = \bar{y}^2 \vee 1 \quad \text{with} \quad \bar{y} = \sup \{y \geq 0 \mid \mathbf{w}^S(\mathbf{s}, \bar{\mathbf{s}}, y) > c_0 y^2\}$$

belongs to the interval  $[1, 2nc_0^{-1}]$ . It follows from this definition of  $D^S$  that

$$(24) \quad \mathbf{w}^S(\mathbf{s}, \bar{\mathbf{s}}, y) \leq c_0 y^2 \quad \text{for all } y > \sqrt{D^S(\mathbf{s}, \bar{\mathbf{s}})}.$$

**Theorem 3.** *Let  $S$  be a countable and identifiable subset of  $\mathcal{L}_0$ . The estimation procedure described in Section 4.2 leads to the following bound which is valid for any  $\rho$ -estimator  $\hat{\mathbf{s}}$  based on  $S$ , all  $\mathbf{s}$  in  $\mathcal{L}_0$  and all  $\xi > 0$  :*

$$(25) \quad \mathbb{P}_{\mathbf{s}} \left[ \mathbf{h}^2(\mathbf{s}, \hat{\mathbf{s}}) \leq \inf_{\bar{\mathbf{s}} \in S} \{c_1 \mathbf{h}^2(\mathbf{s}, \bar{\mathbf{s}}) - \mathbf{h}^2(\mathbf{s}, S) + c_2 D^S(\mathbf{s}, \bar{\mathbf{s}})\} + c_3(1.45 + \xi) \right] \geq 1 - e^{-\xi}.$$

This implies in particular that

$$(26) \quad \mathbb{E}_{\mathbf{s}} [\mathbf{h}^2(\mathbf{s}, \hat{\mathbf{s}})] \leq \inf_{\bar{\mathbf{s}} \in S} \{c_1 \mathbf{h}^2(\mathbf{s}, \bar{\mathbf{s}}) - \mathbf{h}^2(\mathbf{s}, S) + c_2 D^S(\mathbf{s}, \bar{\mathbf{s}})\} + c_4$$

and, more generally,

$$(27) \quad \mathbb{E}_{\mathbf{s}} [\mathbf{h}^\ell(\mathbf{s}, \hat{\mathbf{s}})] \leq C(\ell) \left[ \inf_{\bar{\mathbf{s}} \in S} \left\{ \mathbf{h}^\ell(\mathbf{s}, \bar{\mathbf{s}}) + (D^S(\mathbf{s}, \bar{\mathbf{s}}))^{\ell/2} \right\} \right] \quad \text{for all } \ell \geq 1.$$

The proof will be provided in Section 9.1. If we set

$$(28) \quad D^S = \sup_{(\mathbf{s}, \bar{\mathbf{s}}) \in \mathcal{L}_0 \times S} D^S(\mathbf{s}, \bar{\mathbf{s}}) \quad \text{and} \quad \bar{D}^S = \sup_{(\mathbf{s}, \bar{\mathbf{s}}) \in \mathcal{L}_0 \times \mathcal{L}_0} D^S(\mathbf{s}, \bar{\mathbf{s}}),$$

then (25) becomes

$$(29) \quad \mathbb{P}_{\mathbf{s}} [\mathbf{h}^2(\mathbf{s}, \hat{\mathbf{s}}) \leq (c_1 - 1) \mathbf{h}^2(\mathbf{s}, S) + c_2 D^S + c_3(1.45 + \xi)] \geq 1 - e^{-\xi}.$$

**Remark:** In the sequel, we shall often content ourselves to provide our results in the form of exponential deviations similar to (25) and (29) like

$$\mathbb{P}_{\mathbf{s}} [\mathbf{h}^2(\mathbf{s}, \hat{\mathbf{s}}) \leq \Gamma + c\xi] \geq 1 - e^{-\xi} \quad \text{for all } \xi > 0,$$

where  $\Gamma$  depends on various quantities involved in our assumptions. Such a deviation bound immediately leads by integration to  $\mathbb{E}_{\mathbf{s}} [\mathbf{h}^2(\mathbf{s}, \hat{\mathbf{s}})] \leq \Gamma + c$  and also implies bounds similar to (27) for the moments of  $\mathbf{h}(\mathbf{s}, \hat{\mathbf{s}})$  as well as risk bounds for more general loss functions of the form  $\ell(\mathbf{h}(\mathbf{s}, \hat{\mathbf{s}}))$ .

In the forthcoming sections, we shall use this central Theorem to establish risk bounds for our estimator over more general models than just countable ones. Before turning to these bounds, let us note here that we can already deduce from (26) (taking  $\bar{\mathbf{s}} = \mathbf{s}$ ) that the estimator  $\hat{\mathbf{s}}$  satisfies, since  $D^S(\bar{\mathbf{s}}, \bar{\mathbf{s}}) \geq 1$ ,

$$(30) \quad \sup_{\mathbf{s} \in S} \mathbb{E}_{\mathbf{s}} [\mathbf{h}^2(\mathbf{s}, \hat{\mathbf{s}})] \leq Cd(S) \quad \text{with} \quad d(S) = \sup_{\bar{\mathbf{s}} \in S} D^S(\bar{\mathbf{s}}, \bar{\mathbf{s}}).$$

The assumption that  $\mathbf{s}$  belongs to  $S$  is quite restrictive, nevertheless we shall see in the next section that a bound of this type is not only true for the elements  $\mathbf{s}$  lying in  $S$  but also for those which are close enough to  $S$  with respect to the Kullback-Leibler divergence.

**4.4. Robustness properties with respect to the Kullback-Leibler divergence.** We recall that the Kullback-Leibler divergence (KL-divergence for short) between two probabilities  $P$  and  $Q$  on  $\mathcal{X}$  is given by

$$K(P, Q) = \int \log \left( \frac{dP}{dQ} \right) dP \in [0, +\infty] \quad \text{if } P \ll Q \quad \text{and} \quad K(P, Q) = +\infty \quad \text{otherwise.}$$

For  $\mathbf{t}, \mathbf{t}' \in \mathcal{L}_0$ , we shall set for simplicity

$$\mathbf{K}(\mathbf{t}, \mathbf{t}') = K \left( \bigotimes_{i=1}^n (t_i \cdot \mu_i), \bigotimes_{i=1}^n (t'_i \cdot \mu_i) \right) = \sum_{i=1}^n K(t_i \cdot \mu_i, t'_i \cdot \mu_i).$$

It is well-known that  $2\mathbf{h}^2(\mathbf{t}, \mathbf{t}') \leq \mathbf{K}(\mathbf{t}, \mathbf{t}')$  for all  $\mathbf{t}, \mathbf{t}' \in \mathcal{L}_0$ .

**Theorem 4.** Let  $\overline{S}$  be a model and  $S$  a countable subset of  $\overline{S}$  satisfying

$$(31) \quad \inf_{\overline{\mathbf{s}} \in S} \mathbf{K}(\mathbf{s}, \overline{\mathbf{s}}) = \inf_{\overline{\mathbf{s}} \in \overline{S}} \mathbf{K}(\mathbf{s}, \overline{\mathbf{s}}) = \mathbf{K}(\mathbf{s}, \overline{S}) \quad \text{for all } \mathbf{s} \in \mathcal{L}_0.$$

Then, any  $\rho$ -estimator  $\widehat{\mathbf{s}}$  based on  $S$  satisfies, with  $d(S)$  given by (30),

$$(32) \quad \mathbb{E}_{\mathbf{s}} [\mathbf{h}^2(\mathbf{s}, \widehat{\mathbf{s}})] \leq C [\mathbf{K}(\mathbf{s}, \overline{S}) + d(S)] \quad \text{for all } \mathbf{s} \in \mathcal{L}_0$$

and some universal constant  $C$ .

*Proof.* It relies on Theorem 3 and the following proposition (to be proved in Section 9.3) which is a variant of the lemma (Section 5.3) in Barron (1991) and of independent interest since it applies to many other situations.

**Proposition 5.** Let  $\mathbf{s}$  and  $\overline{\mathbf{s}}$  belong to  $\mathcal{L}_0$  and  $T(\mathbf{X})$  be a random variable such that

$$\mathbb{P}_{\overline{\mathbf{s}}} [T(\mathbf{X}) \geq z] \leq ae^{-z} \quad \text{for all } z \geq 0 \quad \text{and some } a > 0.$$

Then, if  $c = \log(1 + a) + \mathbf{K}(\mathbf{s}, \overline{\mathbf{s}})$ ,

$$(33) \quad \mathbb{E}_{\mathbf{s}} [T(\mathbf{X})] \leq 1 + c + \log(1 + c + \sqrt{2c}) < 1 + c + \sqrt{2c}.$$

If  $\mathbb{P}_{\overline{\mathbf{s}}} [T(\mathbf{X}) \geq z] \leq ae^{-bz}$  for all  $z \geq z_0 \geq 0$  with  $a, b > 0$ , then

$$(34) \quad \mathbb{E}_{\mathbf{s}} [T(\mathbf{X})] \leq z_0 + b^{-1} (1 + c' + \sqrt{2c'}) \quad \text{with } c' = \log(1 + ae^{-bz_0}) + \mathbf{K}(\mathbf{s}, \overline{\mathbf{s}}).$$

It follows from (25) with  $\mathbf{s}$  replaced by  $\overline{\mathbf{s}} \in S$  that

$$(35) \quad \mathbb{P}_{\overline{\mathbf{s}}} [\mathbf{h}^2(\overline{\mathbf{s}}, \widehat{\mathbf{s}}) > c_2 D^S(\overline{\mathbf{s}}, \overline{\mathbf{s}}) + 1.45c_3 + \xi] \leq e^{-\xi/c_3} \quad \text{for all } \xi > 0$$

and we may therefore apply (34) to  $T(\mathbf{X}) = \mathbf{h}^2(\overline{\mathbf{s}}, \widehat{\mathbf{s}}) - c_2 D^S(\overline{\mathbf{s}}, \overline{\mathbf{s}}) - 1.45c_3$  with  $a = 1$ ,  $b = 1/c_3$  and  $z_0 = 0$ . Then  $c' = \log 2 + \mathbf{K}$  with  $\mathbf{K} = \mathbf{K}(\mathbf{s}, \overline{\mathbf{s}})$  and

$$\mathbb{E}_{\mathbf{s}} [\mathbf{h}^2(\overline{\mathbf{s}}, \widehat{\mathbf{s}})] \leq c_2 D^S(\overline{\mathbf{s}}, \overline{\mathbf{s}}) + 1.45c_3 + c_3 (1 + \log 2 + \mathbf{K} + \sqrt{2(\log 2 + \mathbf{K})}).$$

We finally derive from the triangular inequality and (7) with  $\alpha = 1/50$  that

$$\mathbb{E}_{\mathbf{s}} [\mathbf{h}^2(\mathbf{s}, \widehat{\mathbf{s}})] \leq \frac{51}{50} [c_2 D^S(\overline{\mathbf{s}}, \overline{\mathbf{s}}) + c_3 (2.45 + \log 2 + \mathbf{K} + \sqrt{2(\log 2 + \mathbf{K})}) + 50\mathbf{h}^2(\mathbf{s}, \overline{\mathbf{s}})].$$

Since  $\overline{\mathbf{s}}$  is arbitrary in  $S$  and  $\mathbf{h}^2 \leq \mathbf{K}/2$ , it follows that

$$\mathbb{E}_{\mathbf{s}} [\mathbf{h}^2(\mathbf{s}, \widehat{\mathbf{s}})] \leq \frac{51}{50} \inf_{\overline{\mathbf{s}} \in S} \left\{ c_2 D^S(\overline{\mathbf{s}}, \overline{\mathbf{s}}) + c_3 (2.45 + \log 2 + \mathbf{K} + \sqrt{2(\log 2 + \mathbf{K})}) + 25\mathbf{K} \right\},$$

hence by (30) and (31),

$$\mathbb{E}_{\mathbf{s}} [\mathbf{h}^2(\mathbf{s}, \widehat{\mathbf{s}})] \leq C \left[ \inf_{\overline{\mathbf{s}} \in S} \mathbf{K}(\mathbf{s}, \overline{\mathbf{s}}) + d(S) \right] = C [\mathbf{K}(\mathbf{s}, \overline{S}) + d(S)]$$

for some universal constant  $C > 0$ . □

Inequality (32) shows that (30) is not only true when  $\mathbf{s}$  belongs to  $S$  but also when it belongs to  $\overline{S}$  provided that (31) holds and that this risk bound deteriorates by at most the additional term  $\mathbf{K}(\mathbf{s}, \overline{S}) = \inf_{\overline{\mathbf{s}} \in \overline{S}} \mathbf{K}(\mathbf{s}, \overline{\mathbf{s}})$  when  $\mathbf{s}$  does not belong to  $\overline{S}$ . The estimator  $\widehat{\mathbf{s}}$  is therefore robust with respect to the KL-divergence.

Similar results actually hold for any estimator  $\widetilde{\mathbf{s}}$  and any non-negative loss function  $\ell$  such that an analogue of (35) is satisfied, more precisely if

$$\mathbb{P}_{\overline{\mathbf{s}}}[\ell(\overline{\mathbf{s}}, \widehat{\mathbf{s}}) > C(\overline{\mathbf{s}}) + \xi] \leq e^{-b\xi} \quad \text{for all } \xi > 0 \text{ and } \overline{\mathbf{s}} \in S.$$

This indeed implies by (34) that

$$\mathbb{E}_{\mathbf{s}}[\ell(\overline{\mathbf{s}}, \widehat{\mathbf{s}})] \leq C(\overline{\mathbf{s}}) + (C'/b) [1 + \mathbf{K}(\mathbf{s}, \overline{\mathbf{s}})]$$

and, if the loss function  $\ell$  satisfies  $\ell(\mathbf{s}, \mathbf{t}) \leq A[\ell(\mathbf{s}, \mathbf{u}) + \ell(\mathbf{u}, \mathbf{t})]$  for some constant  $A$  and all  $\mathbf{s}, \mathbf{t}, \mathbf{u}$ , then

$$\mathbb{E}_{\mathbf{s}}[\ell(\mathbf{s}, \widehat{\mathbf{s}})] \leq A [C(\overline{\mathbf{s}}) + (C'/b) [1 + \mathbf{K}(\mathbf{s}, \overline{\mathbf{s}})] + \ell(\mathbf{s}, \overline{\mathbf{s}})] \quad \text{for all } \overline{\mathbf{s}} \in S$$

and finally

$$\mathbb{E}_{\mathbf{s}}[\ell(\mathbf{s}, \widehat{\mathbf{s}})] \leq A \sup_{\overline{\mathbf{s}} \in S} C(\overline{\mathbf{s}}) + C_0 [1 + \mathbf{K}(\mathbf{s}, S) + \ell(\mathbf{s}, S)] \quad \text{with} \quad \ell(\mathbf{s}, S) = \inf_{\overline{\mathbf{s}} \in S} \ell(\mathbf{s}, \overline{\mathbf{s}}).$$

This means that, if one allows bias terms depending on KL-divergences, which is often the case in density estimation when one uses likelihood-based methods, one can always assume that the true parameter belongs to the model  $S$  and then extend the result to all  $\mathbf{s}$  satisfying  $\inf_{\overline{\mathbf{s}} \in S} \mathbf{K}(\mathbf{s}, \overline{\mathbf{s}}) < +\infty$ .

**4.5. Robustness properties with respect to the Hellinger distance.** Unfortunately, if  $\mathbf{h}^2(\mathbf{s}, \overline{\mathbf{s}}) \leq \mathbf{K}(\mathbf{s}, \overline{\mathbf{s}})/2$ , the reciprocal  $\mathbf{h}^2(\mathbf{s}, \overline{\mathbf{s}}) \geq c\mathbf{K}(\mathbf{s}, \overline{\mathbf{s}})/2$  for some positive  $c$  is definitely not true in general and we cannot use the previous results to get robustness properties with respect to the Hellinger distance, which is actually a much stronger property. Hopefully, this robustness property is already included in our Theorem 3. We do need this robustness property in order to work with general models, not only countable ones. Since our models, as described in Section 4.1 (for instance the classical sets that are used in Approximation Theory), are typically uncountable, we have to replace them by countable approximations.

We shall therefore apply the following strategy: given a model  $\overline{S}$  for  $\mathbf{s}$  replace it by a countable and identifiable approximating set  $S$  to build our estimator. The natural question at this stage is then: “given  $\overline{S}$ , how to choose  $S$ ?”. First,  $S$  should approximate  $\overline{S}$  within some (typically small)  $\eta$ , according to the following definition.

**Definition 6.** *Given a model  $\overline{S}$  in the pseudometric space  $(\mathcal{L}_0, \mathbf{h})$  and  $\eta \geq 0$ , we say that a countable subset  $S[\eta]$  of  $\mathcal{L}_0$  (not necessarily included in  $\overline{S}$ ) is an  $\eta$ -net for  $\overline{S}$  if  $\sup_{\mathbf{t} \in \overline{S}} \mathbf{h}(\mathbf{t}, S[\eta]) \leq \eta$ . In particular a countable and dense subset of  $\overline{S}$  is a 0-net.*

This immediately leads to the following corollary:

**Corollary 7.** *Let  $S = S[\eta]$  be a countable and identifiable  $\eta$ -net for  $\overline{S}$  and  $\widehat{\mathbf{s}}$  a  $\rho$ -estimator based on  $S$ . Then, for all  $\xi > 0$ ,*

$$(36) \quad \mathbb{P}_{\mathbf{s}} \left[ \mathbf{h}^2(\mathbf{s}, \widehat{\mathbf{s}}) \leq c'_1 \mathbf{h}^2(\mathbf{s}, \overline{S}) + \left( c'_1 \eta^2 + c_2 D^{S[\eta]} \right) + c_3(1.45 + \xi) \right] \geq 1 - e^{-\xi},$$

hence

$$(37) \quad \mathbb{E}_{\mathbf{s}} [\mathbf{h}^2(\mathbf{s}, \widehat{\mathbf{s}})] \leq c'_1 \mathbf{h}^2(\mathbf{s}, \overline{S}) + \left( c'_1 \eta^2 + c_2 D^{S[\eta]} \right) + c_4.$$

*Proof.* The first bound follows from the inequality  $\mathbf{h}^2(\mathbf{s}, S[\eta]) \leq 2\mathbf{h}^2(\mathbf{s}, \bar{S}) + 2\eta^2$  applied to (29) and the second one by integration.  $\square$

We see that (37) corresponds to a decomposition of the risk into the sum of three terms among which only one, namely  $c'_1\eta^2 + c_2D^{S[\eta]}$  depends on the chosen net  $S[\eta]$ . We shall therefore focus our attention on a similar quantity, introducing the two following new notions of dimension.

**Definition 8.** Given a model  $\bar{S}$  in  $\mathcal{L}_0$ , we define its dimension  $D(\bar{S})$  and its uniform dimension  $\bar{D}(\bar{S})$  by

$$(38) \quad D(\bar{S}) = \inf_S \left[ 2c_1 \sup_{\mathbf{u} \in \bar{S}} \mathbf{h}^2(\mathbf{u}, S) + c_2 D^S \right] \quad \text{and} \quad \bar{D}(\bar{S}) = \inf_S \left[ 2c_1 \sup_{\mathbf{u} \in \bar{S}} \mathbf{h}^2(\mathbf{u}, S) + c_2 \bar{D}^S \right],$$

where  $D^S$  and  $\bar{D}^S$  have been defined in (28) and, in both cases, the infimum is taken over all countable and identifiable subsets  $S$  of  $\mathcal{L}_0$ .

It follows from these definitions and the bounds  $1 \leq D^S(\mathbf{s}, \bar{S}) \leq 2nc_0^{-1}$  that

$$(39) \quad 1 < c_2 \leq D(\bar{S}) \leq \bar{D}(\bar{S}) \leq 2n(c_1 + c_2c_0^{-1}) < 144n.$$

If  $\bar{S} \subset \bar{S}'$ , then  $D(\bar{S}) \leq D(\bar{S}')$  and  $\bar{D}(\bar{S}) \leq \bar{D}(\bar{S}')$  which means that  $D$  and  $\bar{D}$  are non-decreasing with respect to the inclusion. These two dimensions have actually different purposes: we shall use  $D(\bar{S})$  to bound the risk of a  $\rho$ -estimator on a given model  $\bar{S}$  while  $\bar{D}(\bar{S})$  will be used for model selection purposes in Section 7.

Since the separability of  $\bar{S}$  implies the existence of  $\eta$ -nets  $S[\eta]$  for  $\bar{S}$  whatever  $\eta \geq 0$ , (38) can be reformulated as

$$(40) \quad D(\bar{S}) = \inf_{\eta \geq 0} \inf_{S[\eta]} \left[ 2c_1\eta^2 + c_2D^{S[\eta]} \right] \quad \text{and} \quad \bar{D}(\bar{S}) = \inf_{\eta \geq 0} \inf_{S[\eta]} \left[ 2c_1\eta^2 + c_2\bar{D}^{S[\eta]} \right],$$

where the infima now run over all possible identifiable  $\eta$ -nets  $S[\eta] \subset \mathcal{L}_0$  for  $\bar{S}$ .

The important property of these dimensions lies in the fact that they allow to replace the model  $\bar{S}$  by a suitable  $\eta$ -net  $S[\eta]$  ( $\eta \geq 0$ ) to which Theorem 3 applies. In particular, choosing  $S[\eta]$  such that

$$2c_1\eta^2 + c_2D^{S[\eta]} \leq D(\bar{S}) + c_3/20,$$

which is always possible in view of (40), we get from (36) and an integration with respect to  $\xi$  the following risk bounds.

**Corollary 9.** Given a model  $\bar{S}$ , there exists a  $\rho$ -estimator  $\hat{\mathbf{s}}$  such that for all  $\mathbf{s} \in \mathcal{L}_0$

$$(41) \quad \mathbb{P}_{\mathbf{s}} [\mathbf{h}^2(\mathbf{s}, \hat{\mathbf{s}}) \leq c'_1\mathbf{h}^2(\mathbf{s}, \bar{S}) + D(\bar{S}) + c_3(1.5 + \xi)] \geq 1 - e^{-\xi} \quad \text{for all } \xi > 0$$

and

$$(42) \quad \mathbb{E}_{\mathbf{s}} [\mathbf{h}^2(\mathbf{s}, \hat{\mathbf{s}})] \leq c'_1\mathbf{h}^2(\mathbf{s}, \bar{S}) + D(\bar{S}) + c_4.$$

The replacement of  $\bar{S}$  by a suitable pair  $(\eta, S[\eta])$  leads to the risk bound (42) depending on  $\bar{S}$  only. This means that, given a model  $\bar{S}$ , the risk of  $\hat{\mathbf{s}}$  breaks down, up to numerical constants, into the bias term  $\mathbf{h}^2(\mathbf{s}, \bar{S})$  which depends on the quality of the approximation of  $\mathbf{s}$  by the model  $\bar{S}$  and the dimensional term  $D(\bar{S})$  which measures in some sense the massiveness of  $\bar{S}$ . In particular, in the i.i.d. case, we get

$$\mathbb{E}_{\mathbf{s}} [h^2(s, \hat{s})] \leq c'_1h^2(s, \bar{S}) + n^{-1} [D(\bar{S}) + c_4],$$

as expected.

**4.5.1. Models which are VC-subgraph classes.** The first situation that we shall consider is about models  $\overline{S}$  such that any countable and identifiable subset  $S$  of  $\overline{S}$  satisfies  $D^S \leq D'$  where  $D'$  only depends on  $\overline{S}$  but not on the choice of the subset  $S$ . In such a case it is natural to choose for  $S$  a countable and dense subset of  $\overline{S}$  which is a 0-net for  $\overline{S}$  so that (37) leads to

$$(43) \quad \mathbb{E}_{\mathbf{s}} [\mathbf{h}^2(\mathbf{s}, \widehat{\mathbf{s}})] \leq c'_1 \mathbf{h}^2(\mathbf{s}, \overline{S}) + c_2 D' + c_4.$$

In order to deal with this situation, we first need to prove an auxiliary result and for this we shall consider an element  $\mathbf{t} = (t_1, \dots, t_n)$  in  $\mathcal{L}_0$  as a real-valued function on  $\overline{\mathcal{X}} = \bigcup_{i=1}^n (\{i\} \times \mathcal{X}_i)$  defined by

$$(44) \quad \mathbf{t}(\overline{x}) = t_i(x) \quad \text{for all } \overline{x} = (i, x) \in \overline{\mathcal{X}}.$$

Replacing the  $X_i$  by the random variables  $\overline{X}_i = (i, X_i)$  so that  $\mathbf{t}(\overline{X}_i) = t_i(X_i)$ , we see that  $\mathbf{w}^S(\mathbf{s}, \overline{\mathbf{s}}, y)$  can be written as

$$(45) \quad \mathbf{w}^S(\mathbf{s}, \overline{\mathbf{s}}, y) = \mathbb{E}_{\mathbf{s}} \left[ \sup_{\mathbf{f} \in \mathcal{F}^S(\mathbf{s}, \overline{\mathbf{s}}, y)} \left| \sum_{i=1}^n (\mathbf{f}(\overline{X}_i) - \mathbb{E}_{\mathbf{s}} [\mathbf{f}(\overline{X}_i)]) \right| \right],$$

where the supremum runs among the class  $\mathcal{F}^S(\mathbf{s}, \overline{\mathbf{s}}, y)$  of real-valued functions  $\mathbf{f}$  on  $\overline{\mathcal{X}}$  given by

$$\mathcal{F}^S(\mathbf{s}, \overline{\mathbf{s}}, y) = \left\{ \psi \left( \sqrt{\mathbf{t}/\overline{\mathbf{s}}} \right) \mid \mathbf{t} \in \mathcal{B}^S(\mathbf{s}, \overline{\mathbf{s}}, y) \right\}.$$

For a set of real-valued functions  $\mathcal{F}$  on  $\overline{\mathcal{X}}$  and a probability  $Q$  on  $\overline{\mathcal{X}}$ , we denote by  $N(\mathcal{F}, Q, \eta)$  the  $\eta$ -covering number of  $\mathcal{F}$  with respect to  $Q$ , that is, the smallest number of closed balls (with respect to the distance in  $\mathbb{L}_2(Q)$ ) with centers in  $\mathcal{F}$  and radius  $\eta$  needed to cover  $\mathcal{F}$ . We finally introduce the following assumption for a function  $\mathcal{H}$ .

**Assumption 10.** *The function  $\overline{\mathcal{H}}$  defined on  $[1/2, +\infty)$  is non-negative, non-decreasing and*

$$L = \sup_{x \geq 1/2} \left\{ x [\overline{\mathcal{H}}(x)]^{-1/2} \int_x^{+\infty} u^{-2} \sqrt{\overline{\mathcal{H}}(u)} du \right\} < +\infty.$$

The next result will be proved in Section 9.4.

**Proposition 11.** *For all  $y > 0$ , assume that there exists a function  $\overline{\mathcal{H}}_y$ , possibly depending on  $y$  and satisfying Assumption 10 with  $L = L_y$ , such that for all  $\mathbf{s}, \overline{\mathbf{s}} \in \mathcal{L}_0$ ,*

$$(46) \quad \log N \left( \mathcal{F}^S(\mathbf{s}, \overline{\mathbf{s}}, y), \frac{1}{n} \sum_{i=1}^n \delta_{\overline{X}_i(\omega)}, z \right) \leq \overline{\mathcal{H}}_y \left( \frac{1}{z} \right) \quad \text{for all } \omega \in \Omega \text{ and } 0 < z \leq 2.$$

There exists a universal constant  $C_0$  such that

$$(47) \quad \mathbf{w}^S(\mathbf{s}, \overline{\mathbf{s}}, y) \leq C_0 \left[ y L_y \sqrt{6 H_y} + L_y^2 H_y \right] \quad \text{with} \quad H_y = \overline{\mathcal{H}}_y \left( \sqrt{\frac{n}{24 y^2}} \vee \frac{1}{2} \right),$$

hence

$$(48) \quad D^S(\mathbf{s}, \overline{\mathbf{s}}) \leq \overline{D}^S \leq \sup \left\{ y^2 \mid 0 < y^2 < \frac{2C_0}{c_0} \left( 1 + \frac{3C_0}{c_0} \right) L_y^2 H_y \right\} \vee 1.$$



It happens that an inequality such as (46) is typically satisfied for VC-subgraph classes  $\mathcal{F}^S$ . In order to avoid a long digression, we defer the relevant definitions, properties and proofs about VC-subgraph classes to Section 8. At this stage, it is sufficient to recall that the index  $\overline{V}$  of a VC-subgraph class is a positive integer. Our main result about models  $\overline{S}$  which are VC-subgraph classes is as follows.

**Theorem 12.** *If  $\overline{S}$ , viewed as a set of real-valued functions on  $\overline{\mathcal{X}}$  as defined by (44), is VC-subgraph with index  $\overline{V}$ , then for all countable, identifiable and dense subsets  $S$  of  $\overline{S}$ ,*

$$(49) \quad \overline{D}(\overline{S}) \leq c_2 \overline{D}^S \leq C \overline{V} [1 + \log_+(n/\overline{V})]$$

for some universal constant  $C$ . Consequently, any  $\rho$ -estimator  $\widehat{\mathbf{s}}$  based on such an  $S$  satisfies, whatever  $\mathbf{s} \in \mathcal{L}_0$ ,

$$(50) \quad \mathbb{P}_{\mathbf{s}} [C' \mathbf{h}^2(\mathbf{s}, \widehat{\mathbf{s}}) \leq \mathbf{h}^2(\mathbf{s}, \overline{S}) + \overline{V} [1 + \log_+(n/\overline{V})] + \xi] \geq 1 - e^{-\xi} \quad \text{for all } \xi > 0$$

and

$$(51) \quad C'' \mathbb{E}_{\mathbf{s}} [\mathbf{h}^2(\mathbf{s}, \widehat{\mathbf{s}})] \leq \mathbf{h}^2(\mathbf{s}, \overline{S}) + \overline{V} [1 + \log_+(n/\overline{V})]$$

for some universal constants  $C', C'' > 0$ . If  $\mathcal{X}_i = \mathcal{X}$  for all  $i$  and  $\overline{S}$  is of the form  $\{\mathbf{t} = (t, \dots, t), t \in \Theta\}$  for a set  $\Theta$  of real valued functions on  $\mathcal{X}$  which is VC-subgraph with index  $\overline{V}$ , the previous bound still holds.

*Proof.* Let  $S$  be any countable subset of  $\overline{S}$ . Then it is VC-subgraph with index not larger than  $\overline{V}$ . Since for all  $\mathbf{s}$  and  $\overline{\mathbf{s}}$  in  $\mathcal{L}_0$  and  $y > 0$ ,  $\mathcal{F}^S(\mathbf{s}, \overline{\mathbf{s}}, y) \subset \left\{ \psi \left( \sqrt{\mathbf{t}/\overline{\mathbf{s}}} \right) \mid \mathbf{t} \in S \right\}$ , it follows from (vii) of Proposition 42 that  $\mathcal{F}^S(\mathbf{s}, \overline{\mathbf{s}}, y)$  is VC-subgraph with index not larger than  $\overline{V}$ . Then, by (94) below, there exists a universal constant  $A$  such that, for all  $\mathbf{s}, \overline{\mathbf{s}} \in \mathcal{L}_0$ ,  $y, z > 0$  and any probability  $Q$  on  $\overline{\mathcal{X}}$ ,

$$(52) \quad \log N(\mathcal{F}^S(\mathbf{s}, \overline{\mathbf{s}}, y), Q, z) \leq 2\overline{V} \log_+(A/z).$$

Proposition 11 therefore applies with  $\overline{\mathcal{H}}_y(x) = 2\overline{V} \log_+(Ax)$  and we may assume that  $A \geq 2e$  so that  $\log_+(Au) = \log(Au) \geq 1$  for  $u \geq 1/2$ . An integration by parts then leads, for  $x \geq 1/2$ , to

$$\frac{\int_x^{+\infty} u^{-2} \sqrt{\log(Au)} du}{x^{-1} \sqrt{\log(Ax)}} = 1 + \frac{x}{\sqrt{\log(Ax)}} \int_x^{+\infty} \frac{du}{2u^2 \sqrt{\log(Au)}} < 1 + x \int_x^{+\infty} \frac{du}{2u^2} = \frac{3}{2},$$

which shows that  $L_y \leq 3/2$  and (49) follows from (48). Inequalities (50) and (51) derive from (36) and (37) respectively with  $S[\eta] = S$  and  $\eta = 0$  since  $\overline{V} \geq 1$ .  $\square$

**4.5.2. Models which are totally bounded.** Of course, not all models are VC-subgraph classes but there exists another type of models for which we are able to bound  $D^{S[\eta]}$  for suitable  $\eta$ -nets of  $\overline{S}$ . When  $\overline{S}$  is totally bounded, one can take  $S[\eta]$  finite for all  $\eta > 0$  and so are the subsets  $\mathcal{B}^{S[\eta]}(\mathbf{s}, \overline{\mathbf{s}}, y)$  of  $S[\eta]$  for all positive  $y$  and  $\eta$ . Conversely, if, for all  $y, \eta > 0$ , one can choose  $S[\eta]$  so that the sets  $\mathcal{B}^{S[\eta]}(\mathbf{s}, \overline{\mathbf{s}}, y)$  are finite, this is in particular true for  $y = \sqrt{2n}$  and, since the distance  $\mathbf{h}$  is bounded by  $\sqrt{n}$ ,  $S[\eta] = \mathcal{B}^{S[\eta]}(\mathbf{s}, \overline{\mathbf{s}}, \sqrt{2n})$  is finite for all  $\eta > 0$ . This implies that  $\overline{S}$  is totally bounded so that this approach based on the cardinality of  $\mathcal{B}^{S[\eta]}(\mathbf{s}, \overline{\mathbf{s}}, y)$  is restricted to totally bounded models only. It nevertheless has the advantage to require the control of the supremum of the process  $|\mathbf{Z}(\mathbf{X}, \overline{\mathbf{s}}, \cdot)|$  over a finite set which can be done via the following result to be proved in Section 9.5.

**Proposition 13.** *Let  $\mathbf{s}$  and  $\bar{\mathbf{s}}$  belong to  $\mathcal{L}_0$ ,  $y > 0$  and  $S$  be a countable and identifiable subset of  $\mathcal{L}_0$ . Assume that  $|\mathcal{B}^S(\mathbf{s}, y)| < +\infty$ , then*

$$\mathbf{w}^S(\mathbf{s}, \bar{\mathbf{s}}, y) \leq 2 \left[ y \sqrt{3 \log_+(N)} + \log_+(N) \right] \quad \text{with} \quad N = 2 |\mathcal{B}^S(\mathbf{s}, y)|.$$

With such a result at hand, bounding  $D^{S[\eta]}(\mathbf{s}, \bar{\mathbf{s}})$  amounts to controlling  $|S[\eta] \cap \mathcal{B}(\mathbf{s}, y)|$  when  $S[\eta]$  is minimal. Since  $\mathbf{s}$  is unknown, we need to bound the number of points of  $S[\eta]$  lying in an arbitrary Hellinger ball of radius  $y$ . It is then natural to introduce the following entropy bounds.

**Definition 14.** *Given a totally bounded model  $\bar{S}$  of  $\mathcal{L}_0$ ,  $\eta > 0$  and  $S[\eta]$  an  $\eta$ -net for  $\bar{S}$ , we set*

$$\mathcal{H}^{\bar{S}}(\eta, S[\eta], y) = \sup_{\mathbf{s} \in \mathcal{L}_0} \log |S[\eta] \cap \mathcal{B}(\mathbf{s}, y)| \geq 0 \quad \text{for } y \geq \eta.$$

We shall say that  $\bar{S}$  has an entropy dimension bounded by  $V \geq 0$  if, for all  $\eta > 0$ , there exists some  $\eta$ -net  $S[\eta]$  for  $\bar{S}$  such that

$$(53) \quad \mathcal{H}^{\bar{S}}(\eta, S[\eta], y) \leq V \log(y/\eta) \quad \text{for all } y \geq 2\eta.$$

Let  $\tilde{D}$  be a right-continuous function from  $(0, +\infty)$  into  $[1/2, +\infty]$  with  $\tilde{D}(\eta) = 1/2$  for  $\eta \geq \sqrt{n}$ . We shall say that  $\bar{S}$  has a metric dimension bounded by  $\tilde{D}(\cdot)$  if, for all  $\eta > 0$ , there exists some  $\eta$ -net  $S[\eta]$  for  $\bar{S}$  such that

$$(54) \quad \mathcal{H}^{\bar{S}}(\eta, S[\eta], y) \leq (y/\eta)^2 \tilde{D}(\eta) \quad \text{for all } y \geq 2\eta.$$

The definition of the metric dimension is due to Birgé (2006) (Definition 6 p. 293). Since the distance  $\mathbf{h}$  that we use here is bounded by  $\sqrt{n}$ , any singleton  $\{\mathbf{t}\}$  in  $\mathcal{L}_0$  is a  $\sqrt{n}$ -net for any subset of  $\mathcal{L}_0$  so that  $\mathcal{H}^{\bar{S}}(\eta, \{\mathbf{t}\}, y) = 0$  for  $y/2 \geq \eta \geq \sqrt{n}$  and we can always set  $\tilde{D}(\eta) = 1/2$  for  $\eta \geq \sqrt{n}$ . The logarithm being a slowly varying function, it is not difficult to see that the notion of metric dimension is more general than the entropy one in the sense that if  $\bar{S}$  has an entropy dimension bounded by some  $V$ , then it also has a metric dimension bounded by  $\tilde{D}(\cdot)$  with

$$(55) \quad \tilde{D}(\eta) \leq (1/2) \vee [V(\log 2)/4] \quad \text{for all } \eta > 0.$$

**Proposition 15.** *Let  $\bar{S}$  be a totally bounded nonempty subset of  $\mathcal{L}_0$  with metric dimension bounded by  $\tilde{D}(\cdot)$ . Let  $\bar{\eta}$  be defined by*

$$\bar{\eta} = \inf \left\{ \eta > 0 \mid \eta^{-2} \tilde{D}(\eta) \leq 8c_0^2/131 \right\}.$$

*Then one can find an  $\bar{\eta}$ -net  $S[\bar{\eta}]$  for  $\bar{S}$  which satisfies  $\overline{D}^{S[\bar{\eta}]} \leq 4\bar{\eta}^2$ . Hence  $\overline{D}(\bar{S}) \leq 2(c_1 + 2c_2)\bar{\eta}^2$  and any  $\rho$ -estimator  $\hat{\mathbf{s}}$  based on  $S[\bar{\eta}]$  satisfies*

$$\mathbb{P}_{\mathbf{s}} [\mathbf{h}^2(\mathbf{s}, \hat{\mathbf{s}}) \leq c'_1 \mathbf{h}^2(\mathbf{s}, \bar{S}) + 2(c_1 + 2c_2)\bar{\eta}^2 + c_3(1.45 + \xi)] \geq 1 - e^{-\xi} \quad \text{for all } \xi > 0.$$

*Proof.* For  $\eta > 0$ , let  $S[\eta]$  be a minimal  $\eta$ -net for  $\bar{S}$ . Using (54) and the fact that  $\tilde{D}(\eta) \geq 1/2$ , we derive that

$$\log_+ \left( 2 |\mathcal{B}^{S[\eta]}(\mathbf{s}, y)| \right) \leq \log 2 + \frac{y^2 \tilde{D}(\eta)}{\eta^2} \leq \left( 1 + \frac{\log 2}{2} \right) \frac{y^2 \tilde{D}(\eta)}{\eta^2} \quad \text{for all } y \geq 2\eta.$$

If, moreover,  $y \geq 2\bar{\eta}$ , using the fact that  $\tilde{D}(\eta)$  is right-continuous and the definition of  $\bar{\eta}$ , we see that

$$\frac{y^2 \tilde{D}(\bar{\eta})}{\bar{\eta}^2} \leq \frac{8c_0^2 y^2}{131},$$

so that we can apply Proposition 13 with

$$\log_+ \left( 2 |\mathcal{B}^{S[\bar{\eta}]}(\mathbf{s}, y)| \right) \leq [1 + (\log 2)/2] [8c_0^2 y^2 / 131] = ay^2$$

and get

$$\mathbf{w}^{S[\bar{\eta}]}(\mathbf{s}, \bar{\mathbf{s}}, y) \leq 2 \left( a + \sqrt{3a} \right) y^2 < c_0 y^2 \quad \text{for all } \mathbf{s}, \bar{\mathbf{s}} \in \mathcal{L}_0 \text{ and } y \geq 2\bar{\eta}.$$

Therefore  $D^{S[\bar{\eta}]}(\mathbf{s}, \bar{\mathbf{s}}) \leq 4\bar{\eta}^2$  which leads to the bounds for  $\overline{D}^{S[\bar{\eta}]}$  and  $\overline{D}(\bar{S})$ . The bound for  $\mathbf{h}^2(\mathbf{s}, \hat{\mathbf{s}})$  then follows from (36).  $\square$

**Remark:** Since  $\tilde{D}(\bar{\eta}) \geq 1/2$ ,  $2(c_1 + 2c_2)\bar{\eta}^2 \geq (131/8)(c_1 + 2c_2)c_0^{-2}$ . It follows that the bounds provided by Proposition 15 are trivial if  $n$  is not larger than this last quantity.

**4.5.3. Minimax risk on a model.** Let us now focus on the specific case of the risk of  $\rho$ -estimators over a model  $\bar{S}$  when  $\mathbf{s}$  is an arbitrary point in  $\bar{S}$  or equivalently on the maximal risk of  $\rho$ -estimators over a model  $\bar{S}$  in  $\mathcal{L}_0$  since it provides an upper bound for the minimax risk  $R_M(\bar{S})$  over  $\bar{S}$  defined by

$$R_M(\bar{S}) = \inf_{\tilde{\mathbf{s}}} \sup_{\mathbf{s} \in \bar{S}} \mathbb{E}_{\mathbf{s}} [\mathbf{h}^2(\mathbf{s}, \tilde{\mathbf{s}})],$$

where the infimum runs among all possible estimators  $\tilde{\mathbf{s}}$  of  $\mathbf{s}$ . In particular,  $R_M(\bar{S}) \leq \mathcal{R}_\rho(\bar{S})$  where  $\mathcal{R}_\rho(\bar{S}) = \sup_{\mathbf{s} \in \bar{S}} \mathbb{E}_{\mathbf{s}} [\mathbf{h}^2(\mathbf{s}, \hat{\mathbf{s}})]$  denotes the maximal risk of any  $\rho$ -estimator  $\hat{\mathbf{s}}$  over  $\bar{S}$ . Restricting ourselves to  $\rho$ -estimators that satisfy (42) and using (39), we get

$$R_M(\bar{S}) \leq \mathcal{R}_\rho(\bar{S}) \leq D(\bar{S}) + c_4 \leq \overline{D}(\bar{S}) + c_4.$$

It follows that it suffices to bound  $D(\bar{S})$  (or  $\overline{D}(\bar{S})$ ) from above in order to control the minimax risk over  $\bar{S}$  which can be done by using the bounds of the previous sections and results in the next corollary.

**Corollary 16.** *If  $\bar{S}$ , viewed as a set of real-valued functions on  $\overline{\mathcal{X}}$  as defined by (44) is VC-subgraph with index  $\overline{V}$ , then*

$$(56) \quad R_M(\bar{S}) \leq \mathcal{R}_\rho(\bar{S}) \leq c_4 + C\overline{V} [1 + \log_+ (n/\overline{V})].$$

*If  $\bar{S}$  is a totally bounded nonempty subset of  $\mathcal{L}_0$  with metric dimension bounded by  $\tilde{D}(\cdot)$ , then*

$$R_M(\bar{S}) \leq \mathcal{R}_\rho(\bar{S}) \leq c_4 + 2(c_1 + 2c_2)\bar{\eta}^2 \quad \text{with} \quad \bar{\eta} = \inf \left\{ \eta > 0 \mid \eta^{-2} \tilde{D}(\eta) \leq 8c_0^2/131 \right\}.$$

The bound (56) for  $\mathcal{R}_\rho(\bar{S})$  that we derived from Theorem 12 involves a logarithmic factor while one would rather expect a bound of the form  $\mathcal{R}_\rho(\bar{S}) \leq C\overline{V}$ . If we compare this result to (53) (with  $\eta = z\sqrt{n}$  in order to make  $\mathbf{h}$  and the  $\mathbb{L}_2(Q)$ -distance comparable), we see that this phenomenon is due to the entropy bound (52) which is uniform with respect to  $y$ . An entropy bound of the form

$$(57) \quad \log N(\mathcal{F}^S(\mathbf{s}, \bar{\mathbf{s}}, y), Q, z) \leq C\overline{V} \log_+ \left( \frac{Ay/\sqrt{n}}{z} \right) \quad \text{for all } y \text{ and } z > 0$$

would lead to the expected inequality  $\mathcal{R}_\rho(\bar{S}) \leq C\bar{V}$ . Unfortunately, we do not know whether a bound such as (57) is true or not but there exists at least one situation where this extra logarithmic factor can be removed: when  $\bar{S}$  consists of piecewise constant functions. For the sake of simplicity we shall only consider the density framework described in Section 2.3.

**Histograms:** Assume that we are in the density framework and have at hand some countable partition  $\mathcal{I}$  of  $\mathcal{X}$  such that  $0 < \mu(I) < +\infty$  for all  $I \in \mathcal{I}$ . We consider the set  $\bar{S} = \bar{S}_{\mathcal{I}}$  of all densities on  $(\mathcal{X}, \mathcal{A}, \mu)$  which are piecewise constant on each element  $I$  of  $\mathcal{I}$ , which means that

$$(58) \quad \bar{S}_{\mathcal{I}} = \left\{ t = \sum_{I \in \mathcal{I}} \frac{t_I}{\mu(I)} \mathbb{1}_I \mid t_I \geq 0 \text{ for all } I \in \mathcal{I} \text{ and } \sum_{I \in \mathcal{I}} t_I = 1 \right\}.$$

If we choose for  $S$  a subset of  $\bar{S}_{\mathcal{I}}$ , the resulting  $\rho$ -estimator  $\hat{s}$  will therefore be an histogram-type estimator and the following result, to be proved in Section 9.6, holds.

**Proposition 17.** *Let  $S$  be a countable subset of  $\bar{S}_{\mathcal{I}}$  and  $s$  be a density with respect to  $\mu$  such that  $\mathcal{J}(s) = \{I \in \mathcal{I} \mid \int_I s d\mu > 0\}$  is finite. Then, for any  $\bar{s} \in S$ ,  $D^S(s, \bar{s}) \leq 6c_0^{-2} |\mathcal{J}(s)|$ .*

There are various potential applications of this result but let us focus here on the case of a finite measure  $\mu$  and a finite partition  $\mathcal{I}$  so that  $|\mathcal{J}(s)| \leq |\mathcal{I}|$  for all  $s \in \mathcal{L}_0$ . It then follows from the previous proposition and (28) that  $D^S \leq 6c_0^{-2} |\mathcal{I}|$  for all countable subsets  $S$  of  $\bar{S}_{\mathcal{I}}$ , hence  $D(\bar{S}_{\mathcal{I}}) \leq 6c_2c_0^{-2} |\mathcal{I}|$ . In this case  $\bar{S}_{\mathcal{I}}$  is a subset of a linear space with dimension  $|\mathcal{I}|$  and is therefore VC-subgraph with index not larger than  $|\mathcal{I}| + 2$ . Comparing our bound for  $D(\bar{S}_{\mathcal{I}})$  with the one provided by Theorem 12 for  $\bar{D}(\bar{S}_{\mathcal{I}})$  we see that the extra logarithmic factor has disappeared. Nevertheless Proposition 17 only provides an upper bound for  $D(\bar{S}_{\mathcal{I}})$  and not for  $\bar{D}(\bar{S}_{\mathcal{I}})$ .

## 5. CONNECTION WITH THE MAXIMUM LIKELIHOOD ESTIMATOR

Throughout this section, we consider the problem of density estimation from  $n$  i.i.d. observations  $X_1, \dots, X_n$  as described in Section 2.3. Our aim is to show that  $\rho$ -estimation may recover the classical MLE in various situations.

**5.1. Regular parametric models.** We consider here a parametric set of densities  $\{t_\theta, \theta \in \Theta'\}$  on the measured space  $(\mathcal{X}, \mathcal{A}, \mu)$  indexed by some open subset  $\Theta'$  of  $\mathbb{R}^d$  and such that the mapping  $\theta \mapsto P_\theta = t_\theta \cdot \mu$  is one-to-one. Our model is  $\bar{S} = \{t_\theta, \theta \in \Theta\}$  for some  $\Theta \subset \Theta'$  and we set  $\|t\|_\infty = \sup_{x \in \mathcal{X}} |t(x)|$  for any function  $t$  on  $\mathcal{X}$ . There have been a number of different assumptions for the “regularity” of a parametric set of densities. Here we mean a modern version of the notion, as inspired by the pioneering works of Le Cam (1970) and Hájek (1972). One may, for instance, use the definition given in Chapter I, Section 7.1 of Ibragimov and Has'minskiĭ (1981).

**Assumption 18.**

- (i) *The parameter set  $\Theta$  is a compact and convex subset of  $\Theta'$  and the true density  $s$  is an element  $t_\vartheta \in \bar{S}$  such that  $\vartheta$  is an interior point of  $\Theta$ .*
- (ii) *The parametric family  $\{t_\theta, \theta \in \Theta'\}$  is regular and the Fisher Information matrix is invertible on  $\Theta$ .*

(iii) There exists a constant  $A_1$  such that

$$\left\| \sqrt{\frac{t_\theta}{t_{\theta'}}} - \sqrt{\frac{t_{\bar{\theta}}}{t_{\theta'}}} \right\|_\infty \leq A_1 |\bar{\theta} - \theta| \quad \text{for all } \theta, \bar{\theta} \text{ and } \theta' \in \Theta.$$

(iv) With probability tending to one when  $n$  goes to infinity, there exists a maximum likelihood estimator  $\tilde{\theta}_n$  which is consistent.

One can then prove (in Section 9.7):

**Theorem 19.** Let  $\bar{S}$  be a parametric model of densities satisfying Assumption 18 and  $S$  an arbitrary countable and dense subset of  $\bar{S}$ . With probability tending to 1 as  $n$  tends to infinity,  $t_{\tilde{\theta}_n}$  belongs to  $\text{Cl}(\mathcal{E}(\mathbf{X}, S))$  and is therefore a  $\rho$ -estimator.

This result shows that when the model is regular enough and contains the true density,  $\rho$ -estimation allows to recover the MLE, at least when  $n$  is large enough. The numerical study of Mathieu Sart (2016) on very simple statistical models  $\bar{S}$  seems to indicate that our procedure allows to recover the MLE in almost all simulations even when the number of observations  $n$  is small. Consequently, there seems to be some space for improvement in Theorem 19. At least, as we shall see in the next section, Assumption 18 could be weakened.

**5.2. A direct computation on a non-regular model.** In this section, we give an example of a non-regular statistical model (in the usual statistical sense) on which we also recover the MLE with probability 1. This means that the connections between the MLE and  $\rho$ -estimators are not restricted to situations where the parameter is estimated at the usual parametric rate  $n^{-1/2}$ .

Let us consider the problem of estimating  $\theta$  from the observation of a sample  $X_1, \dots, X_n$  of an unknown density  $s$  belonging to the model  $\bar{S} = \{q_\theta = \mathbb{1}_{[-1/2+\theta, 1/2+\theta]}, \theta \in \mathbb{R}\}$ . Elementary calculations show that

$$(59) \quad h^2(q_\theta, q_{\theta'}) = |\theta - \theta'| \wedge 1 \quad \text{for all } \theta, \theta' \in \mathbb{R},$$

hence  $S = \{q_\theta, \theta \in \mathbb{Q}\}$  provides a countable and dense subset of  $\bar{S}$ .

**Proposition 20.** Assume that  $s \in \bar{S}$  and let  $X_{(1)} < \dots < X_{(n)}$  be the order statistics corresponding to our sample. The estimator  $\tilde{\theta}_n = (X_{(1)} + X_{(n)})/2$  of  $\theta$  maximizes the likelihood and  $q_{\tilde{\theta}_n}$  is a  $\rho$ -estimator of  $s$ .

*Proof.* The fact that the likelihood  $\theta \mapsto \prod_{i=1}^n \mathbb{1}_{[X_i-1/2, X_i+1/2]}(\theta)$  is maximal for  $\theta = \tilde{\theta}_n$  is easy to check. It remains to show that if  $s \in \bar{S}$ ,  $q_{\tilde{\theta}_n}$  belongs to  $\text{Cl}(\mathcal{E}(\mathbf{X}, S))$  with probability 1.

In the sequel, we only consider points  $\theta$  and  $\theta'$  that belong to  $\mathbb{Q}$ . Since the density  $q_0$  is even,  $\rho(q_\theta, (q_{\theta'} + q_\theta)/2) = \rho(q_{\theta'}, (q_{\theta'} + q_\theta)/2)$  and therefore

$$\mathbf{T}(\mathbf{X}, \mathbf{q}_\theta, \mathbf{q}_{\theta'}) = \frac{1}{\sqrt{2}} \sum_{i=1}^n \psi \left( \sqrt{\frac{q_{\theta'}}{q_\theta}}(X_i) \right) \quad \text{for all } \theta, \theta' \in \mathbb{Q}.$$

For all  $\theta'$ ,  $q_{\theta'}(\cdot)$  takes its values in  $\{0, 1\}$  and for all  $i \in \{1, \dots, n\}$ ,  $q_{\theta'}(X_i) = 1$  if and only if  $\theta' \in [X_i - 1/2, X_i + 1/2]$ . It follows that  $\theta' \in \mathbb{Q}$  and  $q_{\theta'}(X_i) = 1$  for all  $i$  if and only if  $\theta' \in \hat{\Theta}$  with

$$\hat{\Theta} = [X_{(n)} - 1/2, X_{(1)} + 1/2] \cap \mathbb{Q}.$$

This random subset of  $\mathbb{Q}$  is non-void since, when  $s \in \overline{S}$ , the diameter of  $\widehat{\Theta}$  is  $\Delta(\mathbf{X}) = 1 - (X_{(n)} - X_{(1)}) > 0$   $\mathbb{P}_s$ -a.s. For all  $\theta' \in \widehat{\Theta}$  and  $i \in \{1, \dots, n\}$

$$\psi\left(\sqrt{\frac{q_{\theta'}}{q_{\theta}}}(X_i)\right) = \psi\left(\sqrt{\frac{1}{q_{\theta}(X_i)}}\right) = \mathbb{1}_{\{q_{\theta}=0\}}(X_i),$$

which implies that  $\mathbf{T}(\mathbf{X}, \mathbf{q}_{\theta}, \mathbf{q}_{\theta'}) \geq 1/\sqrt{2}$  if  $\theta \notin \widehat{\Theta}$ , hence

$$\Upsilon(S, \mathbf{q}_{\theta}) = \sup_{\theta' \in \mathbb{Q}} \mathbf{T}(\mathbf{X}, \mathbf{q}_{\theta}, \mathbf{q}_{\theta'}) \geq \sup_{\theta' \in \widehat{\Theta}} \mathbf{T}(\mathbf{X}, \mathbf{q}_{\theta}, \mathbf{q}_{\theta'}) \geq \frac{1}{\sqrt{2}} \quad \text{for } \theta \notin \widehat{\Theta}.$$

For  $\theta \in \widehat{\Theta}$ ,  $q_{\theta}(X_i) = 1$  for all  $i$  so that  $q_{\theta'}(X_i) \leq q_{\theta}(X_i)$  for all  $i$  and  $\mathbf{T}(\mathbf{X}, \mathbf{q}_{\theta}, \mathbf{q}_{\theta'}) \leq 0$  whatever  $\theta'$ . It follows that

$$\Upsilon(S, \mathbf{q}_{\theta}) = \sup_{\theta' \in \mathbb{Q}} \mathbf{T}(\mathbf{X}, \mathbf{q}_{\theta}, \mathbf{q}_{\theta'}) = 0.$$

Hence,  $\theta \mapsto \Upsilon(S, \mathbf{q}_{\theta})$  is minimum for the elements  $\theta \in \widehat{\Theta}$  and  $\{\mathbf{q}_{\theta}, \theta \in \widehat{\Theta}\} \subset \mathcal{E}(\mathbf{X}, S)$ . Since  $\tilde{\theta}_n$  belongs to the closure of  $\widehat{\Theta}$  (with respect to the Euclidean distance) and since for any sequence  $(\theta_j)_{j \geq 1}$  converging towards  $\tilde{\theta}_n$ ,  $\mathbf{q}_{\theta_j}$  converges towards  $\mathbf{q}_{\tilde{\theta}_n}$  with respect to the Hellinger distance by (59),  $\mathbf{q}_{\tilde{\theta}_n}$  belongs to  $\text{Cl}(\mathcal{E}(\mathbf{X}, S))$  and is therefore a  $\rho$ -estimator.  $\square$

**5.3. Risk bounds under entropy with bracketing.** Since, for some specific models of densities  $\overline{S}$  that contain the true density  $s$ , the MLE is a  $\rho$ -estimator with probability close to 1, it is natural to wonder how to compare the performance of these two estimators on more general models  $\overline{S}$ , possibly not containing  $s$ . One way to do so is to compare their risk bounds. In the literature, the risk bounds which are established for the MLE usually take the following form

$$(60) \quad C\mathbb{E}_s[h^2(s, \tilde{s})] \leq K(s, \overline{S}) + \tau_n^2 \vee n^{-1},$$

where  $C$  is a positive universal constant and  $K(s, \overline{S}) = \inf_{t \in \overline{S}} K(s, t)$ . As to the number  $\tau_n^2$ , which usually corresponds to the maximal risk over  $\overline{S}$ , it is obtained by solving an equation depending on the bracketing entropy of  $\overline{S}$ . Such a result appears as Theorem 7.11 in Massart (2007). The aim of this section is to establish an analogue of (60) with the same value of  $\tau_n$  for our  $\rho$ -estimator. Our assumptions are similar to those used by Massart with a slight modification (replacing his assumption (M) by (i) below) which corresponds to the fact that we only use countable models.

**Assumption 21.** *There exists a countable subset  $S$  of  $\overline{S}$  with the following properties.*

- (i) *For all densities  $s$  on  $(\mathcal{X}, \mathcal{A}, \mu)$ ,  $K(s, S) = K(s, \overline{S})$ .*
- (ii) *For all  $\sigma > 0$  and  $\overline{s} \in S$  there exists a non-increasing mapping  $z \mapsto \mathcal{H}_{[]}^S(\overline{s}, \sigma, z)$  from  $(0, +\infty)$  into  $(0, +\infty)$  and a family  $\mathcal{I}(\overline{s}, \sigma, z)$  of pairs of non-negative measurable functions on the measured space  $(\mathcal{X}, \mathcal{A}, \mu)$  such that*

$$\log 2 \leq \log |\mathcal{I}(\overline{s}, \sigma, z)| \leq \mathcal{H}_{[]}^S(\overline{s}, \sigma, z) \quad \text{for all } z > 0.$$

*Moreover, for all  $\mathbf{t} \in \mathcal{B}^S(\overline{s}, \sigma\sqrt{n})$  one can find a pair  $(t_L, t_U) \in \mathcal{I}(\overline{s}, \sigma, z)$  such that  $t_L \leq t \leq t_U$  and*

$$\frac{1}{2} \int (\sqrt{t_U} - \sqrt{t_L})^2 d\mu \leq z^2.$$



- (iii) There exists a non-decreasing function  $\phi$  from  $(0, +\infty)$  into  $(0, +\infty)$  such that  $x \mapsto \phi(x)/x$  is non-increasing on  $(0, +\infty)$  and for which

$$\sup_{\bar{s} \in \bar{S}} \int_0^\sigma \sqrt{\mathcal{H}_{[\cdot]}^S(\bar{s}, \sigma, z)} dz \leq \phi(\sigma).$$

From these assumptions, we can derive the following result to be proved in Section 9.8.

**Theorem 22.** Let  $X_1, \dots, X_n$  be an  $n$ -sample with values in  $(\mathcal{X}, \mathcal{A}, \mu)$  and density  $s$  with respect to  $\mu$ . Let  $\bar{S}$  be a model of densities satisfying Assumption 21 and

$$\tau_n = \inf \{ \sigma > 0, \phi(\sigma) \leq \sqrt{n}\sigma^2 \}.$$

Then there exist universal constants  $C, C' > 0$  such that

$$(61) \quad d(S) = \sup_{\bar{s} \in \bar{S}} D^S(\bar{s}, \bar{s}) \leq (Cn\tau_n^2) \vee 1$$

and for any  $\rho$ -estimator  $\hat{s}$  of  $s$

$$(62) \quad C' \mathbb{E}_s [h^2(s, \hat{s})] \leq K(s, \bar{S}) + \tau_n^2 \vee n^{-1}.$$

**5.4. Histogram estimators.** Let us go back to the framework that we introduced at the end of Section 4.5.3 which means that we consider the problem of estimation of a density  $s$  with respect to  $\mu$  using the model of piecewise constant functions  $\bar{S}_{\mathcal{I}}$  defined by (58) with a countable partition  $\mathcal{I}$  of  $\mathcal{X}$  satisfying  $0 < \mu(I) < +\infty$  for all  $I \in \mathcal{I}$ . Note that this model is identifiable so that  $(\bar{S}_{\mathcal{I}}, h)$  is a metric space. The model  $\bar{S}_{\mathcal{I}}$  can then be identified with the unit simplex  $\mathcal{S}$  in  $[0, 1]^{\mathcal{I}}$  since by (58), for  $t \in \bar{S}_{\mathcal{I}}$ ,  $\sum_{I \in \mathcal{I}} t_I = 1$  with  $t_I = \int_I t(x) d\mu(x)$ . With this identification, the metric space  $(\bar{S}_{\mathcal{I}}, h)$  is topologically equivalent to the separable Euclidean simplex  $\mathcal{S}$  so that  $\bar{S}_{\mathcal{I}}$  is also separable for the distance  $h$ . We finally set  $\bar{S} = \{\mathbf{t} = (t, t, \dots, t), t \in \bar{S}_{\mathcal{I}}\} \subset (\bar{S}_{\mathcal{I}})^n$ .

Given  $n$  i.i.d. observations  $X_1, \dots, X_n$  with values in  $\mathcal{X}$  and density  $t \in \bar{S}_{\mathcal{I}}$  with respect to  $\mu$ , we set  $N_I = \sum_{i=1}^n \mathbb{1}_I(X_i)$  for  $I \in \mathcal{I}$ . The vector  $(N_I)_{I \in \mathcal{I}} \in [0, n]^{\mathcal{I}}$  is a multinomial vector with parameter  $(t_I)_{I \in \mathcal{I}}$ , the MLE over  $\mathcal{S}$  is then given by  $\{\hat{t}_I, I \in \mathcal{I}\}$  with  $\hat{t}_I = N_I/n$  and the corresponding density estimator  $\hat{t} = \sum_{I \in \mathcal{I}} (\hat{t}_I/\mu(I)) \mathbb{1}_I$  of  $t$  is the MLE on the model  $\bar{S}_{\mathcal{I}}$ . It is also the histogram estimator of the true density  $s$  with respect to the partition  $\mathcal{I}$  of  $\mathcal{X}$ .

**Proposition 23.** The histogram estimator  $\hat{t} = \sum_{I \in \mathcal{I}} [N_I/(n\mu(I))] \mathbb{1}_I$  is a  $\rho$ -estimator built on the model  $\bar{S}$ .

*Proof.* For  $t, u \in \bar{S}_{\mathcal{I}}$ ,  $\rho(t, u) = \sum_{I \in \mathcal{I}} \sqrt{t_I u_I}$  and  $(du/dt)(x) = u_I/t_I$  for  $x \in I$  with the convention  $0/0 = 1$ . The definition (20) of the function  $\mathbf{T}$  therefore implies that

$$\begin{aligned} \mathbf{T}(\mathbf{X}, \mathbf{t}, \mathbf{u}) &= \frac{n}{2\sqrt{2}} \sum_{I \in \mathcal{I}} \left[ \sqrt{t_I u_I} + u_I^2 - \sqrt{t_I u_I + t_I^2} \right] + \frac{1}{\sqrt{2}} \sum_{I \in \mathcal{I}} \psi \left( \sqrt{\frac{u_I}{t_I}} \right) N_I \\ &= \frac{n}{2\sqrt{2}} \sum_{I \in \mathcal{I}} \left[ \sqrt{t_I + u_I} (\sqrt{u_I} - \sqrt{t_I}) + 2\hat{t}_I \frac{\sqrt{u_I/t_I} - 1}{\sqrt{1 + (u_I/t_I)}} \right]. \end{aligned}$$

It follows, setting  $\mathcal{J} = \{I \in \mathcal{I} \mid \hat{t}_I > 0\} = \{I \in \mathcal{I} \mid N_I > 0\}$ , that

$$\mathbf{T}(\mathbf{X}, \hat{\mathbf{t}}, \mathbf{u}) = \frac{n}{2\sqrt{2}} \left[ \sum_{I \in \mathcal{J}} (3\hat{t}_I + u_I) \frac{\sqrt{u_I/\hat{t}_I} - 1}{\sqrt{1 + (u_I/\hat{t}_I)}} + \sum_{I \in \mathcal{J}^c} u_I \right].$$

Setting  $x_I = u_I/\hat{t}_I$  for  $I \in \mathcal{J}$  and  $\sigma(u) = \sum_{I \in \mathcal{J}^c} u_I$ , we finally get,

$$\mathbf{T}(\mathbf{X}, \hat{\mathbf{t}}, \mathbf{u}) = \frac{n}{2\sqrt{2}} \left[ \sum_{I \in \mathcal{J}} \hat{t}_I G(x_I) + \sigma(u) \right] \quad \text{with} \quad G(x) = \frac{(\sqrt{x} - 1)(3 + x)}{\sqrt{1 + x}} \quad \text{for all } x \geq 0.$$

Since, for all  $x > 0$ ,

$$G'(x) = \frac{2x^2 - x^{3/2} + 3x + \sqrt{x} + 3}{2\sqrt{x}(1+x)^{3/2}} \quad \text{and} \quad G''(x) = \frac{(x^{5/2} - 5x^{3/2} - 9x - 3)\sqrt{1+x}}{4x(1+x)^3\sqrt{x}},$$

we can see that  $G'(1) = \sqrt{2}$  and  $(1-x)(G'(x) - \sqrt{2}) > 0$  for all  $x \neq 1$ . It follows that if  $u' \neq u$  and  $x'_I = u'_I/\hat{t}_I$  for  $I \in \mathcal{J}$  we get

$$(63) \quad \hat{t}_I [G(x_I) - G(x'_I)] < \sqrt{2} (u_I - u'_I) \quad \text{if either } x_I > x'_I \geq 1 \text{ or } x_I < x'_I \leq 1.$$

We now consider two cases.

— If  $\sigma(u) > 0$  there exists some  $u_{I'} > 0$  for  $I' \in \mathcal{J}^c$  and some  $I \in \mathcal{J}$  with  $u_I < \hat{t}_I$ . It is therefore possible to decrease  $u_{I'}$  to  $u_{I'} - \varepsilon \geq 0$  and increase  $u_I$  to  $u_I + \varepsilon \leq \hat{t}_I$  for  $\varepsilon > 0$  small enough which implies for  $\mathbf{T}$  an increase larger than  $n2^{-3/2}[\sqrt{2} - 1]\varepsilon > 0$ . It follows that

$$\mathbf{T}(\mathbf{X}, \hat{\mathbf{t}}, \mathbf{u}) < \sup_{\mathbf{t} \in \bar{S}} \mathbf{T}(\mathbf{X}, \hat{\mathbf{t}}, \mathbf{t}) \quad \text{for all } \mathbf{u} = (u, u, \dots, u) \text{ such that } \sigma(u) > 0.$$

— If  $\sigma(u) = 0$  and  $u \neq \hat{t}$  one can find  $J, J' \in \mathcal{J}$  with  $x_J < 1 < x_{J'}$ , hence, for  $\varepsilon > 0$  small enough,  $x'_J = x_J + \varepsilon/\hat{t}_J \leq 1$  and  $x'_{J'} = x_{J'} - \varepsilon/\hat{t}_{J'} \geq 1$ . For such an  $u$ , we define  $u'$  by  $u'_J = u_J + \varepsilon$ ,  $u'_{J'} = u_{J'} - \varepsilon$  and  $u'_I = u_I$  for all other  $I \in \mathcal{I}$  so that  $\sum_{I \in \mathcal{I}} u'_I = \sum_{I \in \mathcal{I}} u_I = 1$  and  $u'_I \in \mathcal{S}$  as required. It then follows from (63) that

$$\mathbf{T}(\mathbf{X}, \hat{\mathbf{t}}, \mathbf{u}) - \mathbf{T}(\mathbf{X}, \hat{\mathbf{t}}, \mathbf{u}') < (n/2) [u_J - u'_J + u_{J'} - u'_{J'}] = 0.$$

It follows that, for all  $u \neq \hat{t}$ ,  $\mathbf{T}(\mathbf{X}, \hat{\mathbf{t}}, \mathbf{u}) < \sup_{\mathbf{t} \in \bar{S}} \mathbf{T}(\mathbf{X}, \hat{\mathbf{t}}, \mathbf{t})$  and finally

$$\sup_{\mathbf{t} \in \bar{S}} \mathbf{T}(\mathbf{X}, \hat{\mathbf{t}}, \mathbf{t}) = \mathbf{T}(\mathbf{X}, \hat{\mathbf{t}}, \hat{\mathbf{t}}) = 0.$$

Since  $|\mathcal{J}| \leq n$  and the mapping

$$y \mapsto \frac{\sqrt{u_I} - \sqrt{\hat{t}_I + y}}{\sqrt{\hat{t}_I + y + u_I}} (3\hat{t}_I + y + u_I)$$

is continuous at 0 uniformly with respect to  $u_I \in [0, 1]$  for  $\hat{t}_I > 0$ , replacing  $\bar{S}$  by a dense subset  $S$  and  $\hat{\mathbf{t}}$  by a close enough approximation  $\hat{\mathbf{t}}_\varepsilon$  with  $\varepsilon > 0$  and  $\sigma(\hat{\mathbf{t}}_\varepsilon) = 0$  leads to

$$\Upsilon(S, \hat{\mathbf{t}}_\varepsilon) = \sup_{\mathbf{u} \in S} \mathbf{T}(\mathbf{X}, \hat{\mathbf{t}}_\varepsilon, \mathbf{u}) < \varepsilon.$$

Since  $\varepsilon$  is arbitrary, this proves that the MLE  $\hat{\mathbf{t}}$  belongs to  $\text{Cl}(\mathcal{E}(\mathbf{X}, S))$ .  $\square$

This applies in particular to the example of Section 4.5.3. It also applies to the case of the  $X_i$  taking their values in a finite set  $\mathcal{X} = \{a_1, \dots, a_r\}$ ,  $r > 1$  (or even a countable set  $\mathcal{X} = \{a_j, j \in \mathbb{N}\}$ ) and to the estimation of the density  $s$  of the  $X_i$  with respect to the counting measure  $\mu$  on  $\mathcal{X}$ . Then the MLE  $\hat{s}$  over the set  $\bar{S}$  of all densities on  $\mathcal{X}$  is given by  $\hat{s}(a_j) = N_j/n$  with  $N_j = \sum_{i=1}^n \mathbb{1}_{a_j}(X_i)$  for  $1 \leq j \leq r$  (or  $j \in \mathbb{N}$ ) and it is a  $\rho$ -estimator with respect to the model  $\bar{S}$ . In particular, if the  $X_i$  are i.i.d. Bernoulli variables with parameter  $\theta$ , the empirical mean is a  $\rho$ -estimator.

## 6. EXAMPLES

**6.1. Homoscedastic regression with fixed design.** In this section, we consider the statistical framework described in Section 2.4. Our aim is therefore to estimate the function  $\mathbf{f}$  from the observation of the  $X_i$ .

The choice of a model  $\overline{S}_{q,F}$  corresponds here to those of a density  $q$  (with respect to the Lebesgue measure  $\mu$ ) to approximate  $p$  and of a subset  $F$  of  $\mathbb{R}^n$  to approximate  $\mathbf{f}$ . More precisely, given  $q$  and  $F$ , we define the model  $\overline{S}_{q,F}$  as the set of functions from  $\mathbb{R}^n$  to  $\mathbb{R}^n$  given by

$$\overline{S}_{q,F} = \{\mathbf{x} \mapsto \mathbf{q}_{\mathbf{g}}(\mathbf{x}) = (q(x_1 - g_1), \dots, q(x_n - g_n)) \mid \mathbf{g} \in F\},$$

which is clearly identifiable.

**Assumption 24.** *The density  $q$  is unimodal.*

**Theorem 25.** *Let Assumption 24 be satisfied. If  $F$ , viewed as a class of functions on  $\{1, \dots, n\}$ , is VC-subgraph with index  $\overline{V}$ , then*

$$(64) \quad \overline{D}(\overline{S}_{q,F}) \leq C\overline{V} [1 + \log_+(n/\overline{V})]$$

and the estimator  $\widehat{\mathbf{s}} = \mathbf{q}_{\widehat{\mathbf{f}}}$  built in Section 4.2 and based on a countable and dense subset of  $\overline{S}_{q,F}$  satisfies, for any density  $p$  and vector  $\mathbf{f} \in \mathbb{R}^n$ ,

$$(65) \quad \mathbb{P}_{\mathbf{s}} [C\mathbf{h}^2(\mathbf{p}_{\mathbf{f}}, \mathbf{q}_{\widehat{\mathbf{f}}}) \leq \mathbf{h}^2(\mathbf{p}_{\mathbf{f}}, \overline{S}_{q,F}) + \overline{V} [1 + \log_+(n/\overline{V})] + \xi] \geq 1 - e^{-\xi} \quad \text{for all } \xi > 0.$$

*Proof.* A vector  $\mathbf{g} \in F$  and the element  $\mathbf{q}_{\mathbf{g}} \in \overline{S}_{q,F}$  can both be viewed as functions on  $\overline{\mathcal{X}} = \{1, \dots, n\} \times \mathbb{R}$  defined respectively, for  $\overline{x} = (i, x) \in \overline{\mathcal{X}}$ , by  $\mathbf{g}(\overline{x}) = g_i$  and  $\mathbf{q}_{\mathbf{g}}(\overline{x}) = q(x - g_i)$ . Under Assumption 24, it follows from the properties (iii), (i), (vi) of Proposition 42, that  $\overline{S}_{q,F}$  is VC-subgraph with index not larger than  $C'\overline{V}$ . We conclude with Theorem 12.  $\square$

Since this proof relies on the fact that  $\overline{S}_{q,F}$  is VC-subgraph, Assumption 24 can be replaced by “ $q$  is multimodal with no more than  $k$  modes”. Indeed, in this case the set  $\overline{S}_{q,F}$  is still VC-subgraph but with index bounded by  $C'(k)\overline{V}$  as noticed in the remark at the end of Section 8. It follows that the constants  $C$  appearing in (64) and (65) now depend on  $k$ .

There are various ways of applying the previous theorem according to the type of bound we would like to get. Let us first note that, by the triangular inequality,  $\mathbf{h}(\mathbf{p}_{\mathbf{f}}, \mathbf{q}_{\mathbf{g}}) \leq \mathbf{h}(\mathbf{p}_{\mathbf{f}}, \mathbf{p}_{\mathbf{g}}) + \mathbf{h}(\mathbf{p}_{\mathbf{g}}, \mathbf{q}_{\mathbf{g}})$  and, by translation invariance,  $\mathbf{h}^2(\mathbf{p}_{\mathbf{g}}, \mathbf{q}_{\mathbf{g}}) = nh^2(p, q)$  so that  $\mathbf{h}^2(\mathbf{p}_{\mathbf{f}}, \overline{S}_{q,F}) \leq 2\mathbf{h}^2(\mathbf{p}_{\mathbf{f}}, \overline{S}_{p,F}) + 2nh^2(p, q)$ . Therefore (65) implies that, for all  $\xi > 0$ ,

$$\mathbb{P}_{\mathbf{s}} \left[ C\mathbf{h}^2(\mathbf{p}_{\mathbf{f}}, \mathbf{q}_{\widehat{\mathbf{f}}}) \leq nh^2(p, q) + \inf_{\mathbf{g} \in F} \mathbf{h}^2(\mathbf{p}_{\mathbf{f}}, \mathbf{p}_{\mathbf{g}}) + \overline{V} (1 + \log_+(n/\overline{V})) + \xi \right] \geq 1 - e^{-\xi}$$

and, by the same argument,

$$\mathbb{P}_{\mathbf{s}} \left[ C\mathbf{h}^2(\mathbf{q}_{\mathbf{f}}, \mathbf{q}_{\widehat{\mathbf{f}}}) \leq nh^2(p, q) + \inf_{\mathbf{g} \in F} \mathbf{h}^2(\mathbf{q}_{\mathbf{f}}, \mathbf{q}_{\mathbf{g}}) + \overline{V} (1 + \log_+(n/\overline{V})) + \xi \right] \geq 1 - e^{-\xi}.$$

Noticing that  $\mathbf{h}^2(\mathbf{q}_{\mathbf{f}}, \mathbf{q}_{\widehat{\mathbf{f}}}) \leq 2\mathbf{h}^2(\mathbf{q}_{\mathbf{f}}, \mathbf{p}_{\mathbf{f}}) + 2\mathbf{h}^2(\mathbf{p}_{\mathbf{f}}, \mathbf{q}_{\widehat{\mathbf{f}}})$ , we also derive similarly that

$$(66) \quad \mathbb{P}_{\mathbf{s}} \left[ C\mathbf{h}^2(\mathbf{q}_{\mathbf{f}}, \mathbf{q}_{\widehat{\mathbf{f}}}) \leq nh^2(p, q) + \inf_{\mathbf{g} \in F} \mathbf{h}^2(\mathbf{q}_{\mathbf{f}}, \mathbf{q}_{\mathbf{g}}) + \overline{V} (1 + \log_+(n/\overline{V})) + \xi \right] \geq 1 - e^{-\xi}.$$

This last formula provides a risk bound for the estimation of  $\mathbf{f}$  by  $\widehat{\mathbf{f}}$  when the loss function takes the special form  $\ell(\mathbf{f}, \widehat{\mathbf{f}}) = \mathbf{h}^2(\mathbf{q}_{\mathbf{f}}, \mathbf{q}_{\widehat{\mathbf{f}}})$  for a known density  $q$ :

$$C\mathbb{E}_{\mathbf{s}} \left[ \ell(\mathbf{f}, \widehat{\mathbf{f}}) \right] \leq nh^2(p, q) + \inf_{\mathbf{g} \in F} \ell(\mathbf{f}, \mathbf{g}) + \overline{V} (1 + \log_+(n/\overline{V})).$$

If  $p$  is known so that we can set  $q = p$  we find the usual “bias plus variance” risk bound without the  $nh^2(p, q)$  term,  $\overline{V} (1 + \log_+(n/\overline{V}))$  playing here the role of a variance term. This shows that the price to pay for not knowing the density  $p$  of the errors and replacing it by  $q$  is an additional bias term of order  $nh^2(p, q)$ .

To make this last risk bound more precise, we introduce the following definition.

**Definition 26.** *We shall say that a density  $q$  is of order  $\alpha \in (-1, 1]$  if it satisfies*

$$(67) \quad a_q \left[ |u - v|^{1+\alpha} \wedge A_q^{-1} \right] \leq h^2(q_u, q_v) \leq A_q \left[ |u - v|^{1+\alpha} \wedge A_q^{-1} \right] \quad \text{for all } u, v \in \mathbb{R}$$

and some constants  $A_q \geq a_q > 0$  depending on  $q$ .

The reader can find in Ibragimov and Has'minskiĭ (1981) Chapter VI p. 281 some sufficient conditions on the density  $q$  to ensure that (67) holds. For illustration, we present here some examples borrowed from these authors. The density  $q = \mathbb{1}_{[-1/2, 1/2]}$  is of order 0. For  $\alpha \in (-1, 1)$ , the density  $q(x) = [2(1 + \alpha)]^{-1} (1 - |x|)^{\alpha} \mathbb{1}_{[-1, 1]}(x)$  is of order  $\alpha$  and so is  $q(x) = C(\alpha) \exp[-|x|^{\alpha/2}]$  for  $\alpha \in (0, 1)$ . For  $\alpha > 1$ , this latter density is of order 1. If the translation model  $\theta \mapsto q(\cdot - \theta)$  is regular (which, in this case, is equivalent to the fact that  $\sqrt{q}$  is differentiable in quadratic mean), it is of order 1.

Let us now set, for  $\alpha \in (-1, 1]$ ,  $\mathbf{g}, \mathbf{g}' \in \mathbb{R}^n$  and  $G \subset \mathbb{R}^n$ ,

$$(68) \quad d_{1+\alpha}(\mathbf{g}, \mathbf{g}') = \sum_{i=1}^n \left( |g_i - g'_i|^{1+\alpha} \wedge A_q^{-1} \right) \quad \text{and} \quad d_{1+\alpha}(\mathbf{g}, G) = \inf_{\mathbf{g}' \in G} d_{1+\alpha}(\mathbf{g}, \mathbf{g}').$$

Applying Theorem 25 with the bound (66) leads to the following result.

**Corollary 27.** *Let  $F$  be a subset of  $\mathbb{R}^n$  which is VC-subgraph with index  $\overline{V}$  and  $q$  be a density on  $\mathbb{R}$  of order  $\alpha \in (-1, 1]$  which satisfies Assumption 24. Then the estimator  $\widehat{\mathbf{s}} = \mathbf{q}_{\widehat{\mathbf{f}}}$  satisfies, for all  $\mathbf{f} \in \mathbb{R}^n$  and all  $\xi > 0$ ,*

$$(69) \quad \mathbb{P}_{\mathbf{s}} \left[ d_{1+\alpha}(\mathbf{f}, \widehat{\mathbf{f}}) \leq C'(q) (d_{1+\alpha}(\mathbf{f}, F) + nh^2(p, q) + \overline{V} (1 + \log_+(n/\overline{V})) + \xi) \right] \geq 1 - e^{-\xi}.$$

To comment on this result, let us consider the simple example of the shift model for i.i.d. observations. Assume that the  $X_i$  are i.i.d. with common density  $p(\cdot - \theta)$  for some unknown parameter  $\theta \in \mathbb{R}$  but a known density  $p$  that we assume to be of order  $\alpha \in (-1, 1]$  and to satisfy Assumption 24. In this case  $f_i = \theta$  for all  $i$  and it is natural to fix  $q = p$  and consider as a model for  $\mathbf{f}$  the linear span  $F$  of  $(1, \dots, 1)$  in  $\mathbb{R}^n$ . The distance  $d_{1+\alpha}(\mathbf{f}, \mathbf{g})$  between  $\mathbf{f} = (\theta, \dots, \theta)$  and an element  $\mathbf{g} = (\theta', \dots, \theta')$  of  $F$  becomes

$$d_{1+\alpha}(\mathbf{f}, \mathbf{g}) = n \left( |\theta - \theta'|^{1+\alpha} \wedge A_p^{-1} \right)$$

and we can deduce from (69) that our estimator  $\widehat{\mathbf{f}} = (\widehat{\theta}, \dots, \widehat{\theta})$  satisfies, for  $n$  large enough and with a probability close to 1,

$$|\theta - \widehat{\theta}| \leq C(p, \alpha) [(\log n)/n]^{1/(1+\alpha)}.$$

As soon as  $\alpha \in (-1, 1)$ , the rate we get improves on the usual parametric one  $1/\sqrt{n}$  achieved by the classical least-squares estimator (under suitable moment conditions on the

$\varepsilon_i$ ). Though faster, this rate can still be improved by a logarithmic factor, as, in fact, the maximum likelihood estimator can achieve the rate  $n^{-1/(1+\alpha)}$  (we refer to Theorem 6.3 p. 314 of the book by Ibragimov and Has'minskiĭ (1981)). We do not know whether this extra logarithmic factor is due to our techniques or if it is really necessary in order to get a robust estimator (which is not the case for the MLE) with unbounded models. If we were ready to make the additional assumption that the set  $F$  is totally bounded, we could use a T-estimator or a  $\rho$ -estimator based on a finite model and use the arguments of Section 4.5.2 to get a risk which would not involve this extra  $\log n$  factor. But for unbounded sets, we do not know how to avoid the use VC-subgraph classes and it is their introduction that leads to this  $\log n$  factor.

**6.2. Simple linear regression.** As an illustration of the superiority of  $\rho$ -estimators over the least squares method in some regression frameworks, we consider the very simple situation of observations  $Y_i = a + bx_i + \varepsilon_i$ ,  $1 \leq i \leq n$ , that is a simple linear regression, where the errors  $\varepsilon_i$  are i.i.d. with a known unimodal density  $p$  and satisfy  $\mathbb{E}[\varepsilon_i] = 0$  and  $\text{Var}(\varepsilon_i) = 1$ . We moreover assume that  $n = 2r - 1$  with  $r \geq 2$  is odd and  $x_i = n^{-1}[2i - n - 1]$  ( $x_1 = -1 + 1/n, \dots, x_{r-1} = -2/n, x_r = 0, x_{r+1} = 2/n, \dots, x_n = 1 - 1/n$ ) so that  $x_i \in (-1, 1)$  for all  $i = 1, \dots, n$ ,

$$\sum_{i=1}^n x_i = 0 \quad \text{and} \quad \sum_{i=1}^n x_i^2 = \frac{n^2 - 1}{3n}.$$

This corresponds to an affine regression function and to the model  $F = \{\mathbf{f}(a, b) \in \mathbb{R}^n, (a, b) \in \mathbb{R}^2\}$  with  $f_i(a, b) = a + bx_i$  for  $1 \leq i \leq n$ . It is well known that, in this case, the least squares estimator  $(\tilde{a}, \tilde{b})$  of the parameter  $(a, b)$  satisfies

$$\mathbb{E}[(\tilde{a} - a)^2] = \frac{1}{n} \quad \text{and} \quad \mathbb{E}[(\tilde{b} - b)^2] = \frac{3n}{n^2 - 1} > \frac{3}{n}.$$

Let us now assume that the density  $p$  satisfies

$$(70) \quad h^2(p(\cdot - \theta), p(\cdot - \theta')) \geq c[|\theta - \theta'|^\gamma \wedge 1]$$

for some  $\gamma \in (0, 2)$  and  $c > 0$ . The joint density of the observations  $Y_i$  is  $\prod_{i=1}^n p(y_i - a - bx_i) = \prod_{i=1}^n p_{a+bx_i}(y_i)$  and can be estimated by a  $\rho$ -estimator based on the model  $F$ , resulting in the estimated density  $\prod_{i=1}^n p_{\hat{a}+\hat{b}x_i}(y_i)$  and we know that, with large probability,

$$\sum_{i=1}^n h^2(p_{\hat{a}+\hat{b}x_i}, p_{a+bx_i}) \leq B \log n,$$

for some constant  $B$  independent of  $n$ . Then (70) implies, with  $\alpha = \hat{a} - a$ ,  $\beta = \hat{b} - b$ , that

$$\sum_{i=1}^n \left[ |\hat{a} - a + (\hat{b} - b)x_i|^\gamma \wedge 1 \right] = \sum_{i=1}^n [|\alpha + \beta x_i|^\gamma \wedge 1] \leq B \log n / c.$$

Note that if  $\alpha\beta \geq 0$  then  $|\alpha + \beta x_i| = |\alpha| + |\beta|x_i$  for all  $i \geq r$  and if  $\alpha\beta < 0$  then  $|\alpha + \beta x_i| = |\alpha| + |\beta|x_{2r-i}$  for all  $i \leq r$  so that

$$\begin{aligned} B \log n / c &\geq \sum_{i=1}^n [|\alpha + \beta x_i|^\gamma \wedge 1] \geq \sum_{i=r}^{2r-1} [(|\alpha| + |\beta|x_i)^\gamma \wedge 1] \\ &\geq \sum_{i=r}^{2r-1} \max\{|\alpha|^\gamma \wedge 1, (|\beta|^\gamma \wedge 1)x_i^\gamma\} \geq r \max\{(|\alpha|^\gamma \wedge 1), (|\beta|^\gamma \wedge 1)I\}, \end{aligned}$$

with

$$I = \frac{1}{r} \sum_{i=r+1}^{2r-1} x_i^\gamma \geq \frac{n}{2r} \int_0^{1-1/n} u^\gamma du = \frac{(1-1/n)^{\gamma+1}}{(\gamma+1)(1+1/n)} \geq \frac{(2/3)^\gamma}{2(\gamma+1)}.$$

It follows that

$$\max\{(|\alpha|^\gamma \wedge 1), (|\beta|^\gamma \wedge 1)\} \leq C(\gamma)n^{-1} \log n,$$

hence  $\max\{|\alpha|, |\beta|\} < 1$  for  $n$  large enough and finally

$$|\alpha|^2 + |\beta|^2 = |\hat{a} - a|^2 + |\hat{b} - b|^2 = O_P\left((\log n/n)^{2/\gamma}\right),$$

which improves on the estimation by least-squares, at least when  $n$  is large, since  $\gamma < 2$ .

**6.3. Homoscedastic regression with random design.** In this section, we consider the regression framework with random design described in Section 2.5. Least squares or penalized least squares are the classical estimators which are used in this context and many efforts have been made to analyze their performances under suitable conditions on the moments of the errors and the distribution of the design (see Baraud (2002), Audibert and Catoni (2011) and the references therein). Our point of view is different. We shall rather assume that the distribution of the design is completely unknown while the distribution of the errors is approximately known and symmetric, but possibly without moments. Furthermore, while the  $\mathbb{L}_2$ -norm with respect to the law of the design is the usual loss function that is used for analyzing the performance of the least squares, we shall rather stick to Hellinger-type losses. More precisely, we evaluate the performance of an estimator  $\hat{f}$  of  $f$  by the risk

$$\mathbb{E}_s [\mathbf{h}^2(\mathbf{p}_f, \mathbf{p}_{\hat{f}})] = n \mathbb{E}_s \left[ \int_{\mathcal{W}} h^2(p_{\hat{f}}(w, \cdot), p_f(w, \cdot)) d\nu(w) \right],$$

with  $p_g(w, \cdot) = p(\cdot - g(w))$  for all  $g \in \mathcal{F}$ . This actually corresponds to the use of the loss function  $\ell(g, g')$  on  $\mathcal{F}$  with

$$(71) \quad \ell(g, g') = h^2(p_g, p_{g'}) = \int_{\mathcal{W}} h^2(p_g(w, \cdot), p_{g'}(w, \cdot)) d\nu(w).$$

When the density  $p$  is of order  $\alpha \in (-1, 1]$ , as given by Definition 26, one can relate  $\ell$  to some power of a more classical  $\mathbb{L}_{1+\alpha}$ -loss since then, according to (67),

$$a_p \int_{\mathcal{W}} [|g - g'|^{1+\alpha} \wedge A_p^{-1}] d\nu \leq \ell(g, g') \leq A_p \int_{\mathcal{W}} [|g - g'|^{1+\alpha} \wedge A_p^{-1}] d\nu \quad \text{for all } g, g' \in \mathcal{F}.$$

If, moreover, the  $\mathbb{L}_\infty$ -norms of the elements of  $\mathcal{F}$  are uniformly bounded by some number  $b > 0$ , then

$$\frac{1}{[A_p(2b)^{1+\alpha}] \vee 1} |g - g'|^{1+\alpha} \leq |g - g'|^{1+\alpha} \wedge A_p^{-1} \leq |g - g'|^{1+\alpha} \quad \text{for all } g, g' \in \mathcal{F}$$

and  $\ell(g, g')$  becomes of the same order as  $\|g - g'\|_{1+\alpha, \nu}^{1+\alpha} = \int_{\mathcal{W}} |g - g'|^{1+\alpha} d\nu$  since

$$\frac{a_p}{[A_p(2b)^{1+\alpha}] \vee 1} \|g - g'\|_{1+\alpha, \nu}^{1+\alpha} \leq \ell(g, g') \leq A_p \|g - g'\|_{1+\alpha, \nu}^{1+\alpha}.$$

In particular, we recover the usual  $\mathbb{L}_2$ -loss when  $p$  is of order 1 which is the case for the Gaussian, Cauchy and Laplace distributions among others.

To estimate  $f$  we proceed as follows: we choose a candidate density  $q$  for  $p$  which we assume to be symmetric and unimodal and consider a model  $F \subset \mathcal{F}$ , which is VC-subgraph



with index  $\bar{V}(F)$  to approximate  $f$ . To  $F$ , we associate the model of densities (with respect to  $\nu \otimes \mu$ ) given by

$$\bar{S}_F = \{\mathbf{q}_g = (q_g, \dots, q_g) \mid g \in F\} \quad \text{where} \quad q_g(w, y) = q(y - g(w))$$

and estimate the density  $s$  of  $(W, Y)$  from the observation of  $(W_1, Y_1), \dots, (W_n, Y_n)$  building the corresponding  $\rho$ -estimator from a countable and dense subset  $S$  of  $\bar{S}_F$ . We can apply this procedure without knowing  $\nu$  since, under the assumptions that  $q$  is symmetric and  $\mu$  is the Lebesgue measure, for all  $g, g'$  and  $w \in \mathcal{W}$ ,

$$\int_{\mathbb{R}} \sqrt{q_g(w, y)r(w, y)} d\mu(y) = \int_{\mathbb{R}} \sqrt{q_{g'}(w, y)r(w, y)} d\mu(y) \quad \text{with} \quad r = \frac{q_g + q_{g'}}{2},$$

so that by integration with respect to  $\nu$ ,  $\rho(q_g, r) = \rho(q_{g'}, r)$ . Therefore  $\mathbf{T}((\mathbf{W}, \mathbf{Y}), p_g, p_{g'})$  simply becomes

$$\mathbf{T}((\mathbf{W}, \mathbf{Y}), p_g, p_{g'}) = \frac{1}{\sqrt{2}} \sum_{i=1}^n \psi \left( \sqrt{\frac{p_{g'}(W_i, Y_i)}{p_g(W_i, Y_i)}} \right).$$

As in the proof of Theorem 25, it follows from Proposition 42 that, under Assumption 34,  $\{q_g, g \in F\}$  is VC-subgraph with index not larger than  $C'\bar{V}(F)$ . Applying Theorem 12 then leads to the following result.

**Theorem 28.** *If  $q$  is unimodal and symmetric and  $F$  is a model for  $f$  which is VC-subgraph of index  $\bar{V}(F)$  there exists a  $\rho$ -estimator  $\hat{s} = q_{\hat{f}}$  of  $s = p_f$  such that for all  $\xi > 0$ , with probability at least  $1 - e^{-\xi}$ ,*

$$\begin{aligned} Ch^2(p_f, q_{\hat{f}}) &\leq \inf_{g \in F} h^2(p_f, q_g) + \frac{\bar{V}(F)}{n} \left[ 1 + \log_+ \left( \frac{n}{\bar{V}(F)} \right) \right] + \frac{\xi}{n} \\ &\leq 2h^2(p, q) + 2 \inf_{g \in F} h^2(p_f, p_g) + \frac{\bar{V}(F)}{n} \left[ 1 + \log_+ \left( \frac{n}{\bar{V}(F)} \right) \right] + \frac{\xi}{n}. \end{aligned}$$

If, in particular,  $p = q$ , then for all  $\xi > 0$ ,

$$\mathbb{P}_s \left[ C\ell(f, \hat{f}) \leq \inf_{g \in F} \ell(f, g) + \frac{\bar{V}(F)}{n} \left[ 1 + \log_+ \left( \frac{n}{\bar{V}(F)} \right) \right] + \frac{\xi}{n} \right] \geq 1 - e^{-\xi}.$$

If, moreover, (67) holds and  $\max \{ \sup_{g \in F} \|g\|_{\infty}, \|f\|_{\infty} \} \leq b < +\infty$ , then

$$\mathbb{P}_s \left[ C' \|f - \hat{f}\|_{1+\alpha, \nu}^{1+\alpha} \leq \inf_{g \in F} \|f - g\|_{1+\alpha, \nu}^{1+\alpha} + \frac{\bar{V}(F)}{n} \left[ 1 + \log_+ \left( \frac{n}{\bar{V}(F)} \right) \right] + \frac{\xi}{n} \right] \geq 1 - e^{-\xi},$$

for all  $\xi > 0$  and some constant  $C'$  depending only on  $A_p, a_p, b$  and  $\alpha$ .

**6.4. Further examples for regression problems.** Let us recall that we want to estimate the unknown function  $f$  on  $\mathcal{W}$  in both the random design framework where  $X = (W, Y)$  and  $Y = f(W) + \varepsilon$  and the fixed design framework which corresponds, with analogous notations, to  $X = f(w) + \varepsilon$  with  $w \in \mathcal{W} = \{1, \dots, n\}$ .

As we noticed in the previous sections, when dealing with both regression frameworks, when the model  $F$  for the regression function  $f$  is VC-subgraph, the performance of the estimator  $\hat{f}$  depends on the VC-index  $\bar{V}(F)$ . A common practice to design regression models is to choose for  $F$  a  $D$ -dimensional linear space of functions which, according to Section 8, is VC-subgraph with index bounded by  $D + 2$ . This includes the celebrated “linear model” in the fixed design framework when  $F$  is the linear span of  $D$  linearly independent vectors  $\mathbf{g}^1, \dots, \mathbf{g}^D$  in  $\mathbb{R}^n$  or, equivalently, of  $D$  functions  $g^1, \dots, g^D$  on  $\{1, \dots, n\}$ .

Let us, for a moment, focus on this situation of  $F$  being a  $D$ -dimensional linear space. Classical least squares estimators in the fixed design case lead to risk bounds of order  $D/n$  when the errors are Gaussian or, more generally, have a few moments, but fail miserably when they are Cauchy while, as we have seen in Section 6.1,  $\rho$ -estimators provide the same rate of convergence, apart from an extra  $\log(n/D)$  factor, with the loss function

$$d_2(\mathbf{g}, \mathbf{g}') = \frac{1}{n} \sum_{i=1}^n \left( |g_i - g'_i|^2 \wedge A_q^{-1} \right).$$

When the errors have a uniform distribution, we derive a bound of order  $(D/n) \log(n/D)$  for the loss

$$d_1(\mathbf{g}, \mathbf{g}') = \frac{1}{n} \sum_{i=1}^n \left( |g_i - g'_i| \wedge A_q^{-1} \right)$$

while errors with unbounded densities of the form  $q(x) = [2(1 - \beta)]^{-1}(1 - |x|)^{-\beta} \mathbb{1}_{[-1,1]}(x)$  with  $0 < \beta < 1$  lead to the same bound with the loss

$$d_{1-\beta}(\mathbf{g}, \mathbf{g}') = \frac{1}{n} \sum_{i=1}^n \left( |g_i - g'_i|^{1-\beta} \wedge A_q^{-1} \right).$$

Since subsets of VC-subgraph classes are also VC-subgraph one can restrict  $F$  to be a bounded subset of such a linear space and get similar results for the random design situation, according to Theorem 28, with loss functions of the form  $\|f - \hat{f}\|_{2,\nu}^2$ ,  $\|f - \hat{f}\|_{1,\nu}$  and  $\|f - \hat{f}\|_{1-\beta,\nu}^{1-\beta}$  respectively.

An alternative way of building models that still satisfy the assumptions which are needed to apply our results is as follows. We start from a  $D$ -dimensional linear space  $G$  of functions on  $\mathcal{W}$  and consider some monotone function  $\Psi$ . Finally we take for  $F$  the set  $\{\Psi \circ g, g \in G\}$ . It follows from (ii) of Proposition 42 that  $F$  is still VC-subgraph with index not larger than  $D + 2$  and the previous results still holds. We may replace “monotone” by “unimodal” and get similar results according to (vi) of the same proposition. This allows, given  $D$  independent functions  $g^1, \dots, g^D$ , to use for instance models  $F$  of the following forms:

$$\left\{ \exp \left[ \sum_{j=1}^D \beta_j g^j \right], \beta_j \in \mathbb{R} \text{ for } 1 \leq j \leq D \right\} \quad \text{or} \quad \left\{ \left| \sum_{j=1}^D \beta_j g^j \right|, \beta_j \in \mathbb{R} \text{ for } 1 \leq j \leq D \right\},$$

among many other possibilities.

**6.5. A parametric bound over a set with infinite metric dimension.** In this section, we want to show that, unlike T-estimators, the construction and performance of which heavily depend on the metric dimension of the model that is used, our estimator can, in some cases, achieve a parametric rate which is not connected to its metric dimension. The following illustration given for the density framework described in Section 2.3 is borrowed from Birgé (1983) (Section 6).

Let  $\Lambda$  be any nonvoid subset of  $\mathbb{N}$  and  $\Theta = \Lambda^{\mathbb{N} \setminus \{0\}}$ ,  $\mathcal{X} = \bigcup_{j \geq 1} \Lambda^j$  be respectively the sets of infinite and finite sequences with entries in  $\Lambda$ . Note that the set  $\mathcal{X}$  is countable and that we may introduce on  $\mathcal{X}$  the family of probabilities  $\{P_\theta, \theta \in \Theta\}$  given by

$$P_\theta = \sum_{j \geq 1} 2^{-j} \delta_{(\theta_1, \dots, \theta_j)} \quad \text{for all } \theta = (\theta_1, \dots, \theta_k, \dots) \in \Theta.$$

In the sequel, we denote by  $s_\theta$  the density of  $P_\theta$  with respect to the counting measure on  $\mathcal{X}$ , that is  $s_\theta(x) = P_\theta(\{x\})$  for all  $x \in \mathcal{X}$ , and set  $\overline{S} = \{s_\theta, \theta \in \Theta\}$ , which is identifiable. Our aim is to estimate  $s_\theta$  from the observation of a sample  $X_1, \dots, X_n$ .

It will be convenient to define the following operators :  $\ell(x)$  is the length of an element  $x \in \mathcal{X}$ , that is  $\ell(x) = j$  if  $x \in \Lambda^j$ ,  $\pi_j$  is the operator from  $\Theta$  to  $\Lambda^j$  such that  $\pi_j(\theta) = (\theta_1, \dots, \theta_j)$  and  $\pi_{-1}$  is an operator from  $\mathcal{X}$  to  $\Theta$  such that if  $x \in \Lambda^j$ ,  $\pi_j \circ \pi_{-1}(x) = x$  or, equivalently,  $\pi_{\ell(x)} \circ \pi_{-1}(x) = x$  for all  $x \in \mathcal{X}$ . It follows that  $\pi_{-1}(\mathcal{X})$  is a countable subset of  $\Theta$  and  $S = \{s_\theta, \theta \in \pi_{-1}(\mathcal{X})\}$  is a countable subset of  $\overline{S}$ .

Let  $J$  be the mapping from  $\Theta^2$  to  $\mathbb{N} \cup \{+\infty\}$  defined by

$$J(\theta, \theta') = \sup\{j \in \mathbb{N} \mid \theta_k = \theta'_k \text{ for } 1 \leq k \leq j\} \quad \text{with } \sup \mathbb{N} = +\infty, \quad \sup \emptyset = 0.$$

Since  $s_\theta(x) = 2^{-\ell(x)}$  if  $\pi_{\ell(x)}(\theta) = x$  and  $s_\theta(x) = 0$  otherwise, the Hellinger distance between two densities  $s_\theta$  and  $s_{\theta'}$  with  $\theta, \theta' \in \Theta$  is given by

$$(72) \quad h^2(s_\theta, s_{\theta'}) = 1 - \rho(s_\theta, s_{\theta'}) = 1 - \sum_{j=1}^{J(\theta, \theta')} 2^{-j} = 2^{-J(\theta, \theta')} \quad \text{with } \sum_{\emptyset} = 0; \quad 2^{-\infty} = 0.$$

For any  $x \in \mathcal{X}$  with  $\ell(x) = j$  one can find a subset  $\Theta_j$  of  $\Theta$  with  $|\Theta_j| = |\Lambda|$  and such that  $\pi_j(\theta) = x$  for all  $\theta \in \Theta_j$  but all  $\pi_{j+1}(\theta)$  are different. It suffices for that to let  $\theta_{j+1}$  go across all elements of  $\Lambda$  to build the elements of  $\Theta_j$ . As a consequence,  $h^2(\theta, \theta') = 2^{-j}$  for all  $\theta \neq \theta' \in \Theta_j$  and  $\Theta_j$  is included in a closed ball of radius  $2^{-j/2}$ . This shows that the metric dimension of  $\overline{S}$  can be made arbitrarily large or even infinite by playing with the cardinality of  $\Lambda$ .

Though  $\overline{S}$  can be massive, the parameter  $\theta$  is not difficult to estimate. Let us first observe that, if  $\ell(X_i) = j$ , then  $X_i = \pi_j(\theta)$   $P_\theta$ -a.s. Therefore, if  $k = \ell(X_{i_0}) = \sup_{1 \leq i \leq n} \ell(X_i)$  and  $\widehat{\Theta} = \pi_k^{-1}(X_{i_0})$ , for all  $\widehat{\theta} \in \widehat{\Theta}$ ,

$$(73) \quad s_{\widehat{\theta}}(X_i) = 2^{-\ell(X_i)} = s_\theta(X_i), \quad \text{for } i \in \{1, \dots, n\}, \quad P_\theta\text{-a.s.}$$

while for all  $\theta' \notin \widehat{\Theta}$   $s_{\theta'}(X_i) \leq s_\theta(X_i)$  for  $1 \leq i \leq n$  and  $0 = s_{\theta'}(X_{i_0}) < s_\theta(X_{i_0})$ . It follows that the likelihood reaches its maximum over the elements  $\widehat{\theta} \in \widehat{\Theta}$   $P_\theta$ -a.s. It is proven in Birgé (1983) that these maximum likelihood estimators satisfy

$$(74) \quad \mathbb{E}_\theta [h^2(s_\theta, s_{\widehat{\theta}})] \leq Cn^{-1}, \quad \text{for all } \theta \in \Theta$$

and some numerical constant  $C > 0$ . In particular, the minimax risk over  $\overline{S}$  converges to zero with parametric rate.

Let us now consider the set of  $\rho$ -estimators of  $s_\theta$  build from the model  $S$ , that is the set  $\text{Cl}(\mathcal{E}(\mathbf{X}, S))$ . The following result then holds.

**Proposition 29.** *For all  $\theta \in \Theta$  and any choice of a maximum likelihood estimator  $\widehat{\theta}$  in  $\widehat{\Theta}$ ,  $s_{\widehat{\theta}}$  is a  $\rho$ -estimator of  $s_\theta$ ,  $P_\theta$ -a.s.*

*Proof.* Let us observe that if  $\theta \in \Theta$ ,  $x = \pi_j(\theta)$  and  $\theta' = \pi_{-1}(x)$ , it follows from (72) that  $h^2(s_\theta, s_{\theta'}) \leq 2^{-j}$  so that  $S$  is a dense subset of  $\overline{S}$ . To show that  $\{s_{\widehat{\theta}}, \widehat{\theta} \in \widehat{\Theta}\} \subset \text{Cl}(\mathcal{E}(\mathbf{X}, S))$  it is therefore enough to prove that the elements of  $\widehat{S} = \{s_{\widehat{\theta}}, \widehat{\theta} \in \widehat{\Theta}\} \cap S$  minimize  $\Upsilon(S, \cdot)$

over  $S$ . First note that for  $\theta, \theta' \in \Theta$  with  $\theta \neq \theta'$ ,

$$\begin{aligned} \rho\left(s_\theta, \frac{s_\theta + s_{\theta'}}{2}\right) &= \sum_{x \in \mathcal{X}} \sqrt{s_\theta(x) \frac{s_\theta(x) + s_{\theta'}(x)}{2}} = \sum_{j=1}^{J(\theta, \theta')} 2^{-j} + \sum_{j > J(\theta, \theta')} \frac{2^{-j}}{\sqrt{2}} \\ &= 1 - \left(1 - \frac{1}{\sqrt{2}}\right) 2^{-J(\theta, \theta')} = \rho\left(s_{\theta'}, \frac{s_\theta + s_{\theta'}}{2}\right). \end{aligned}$$

Let us now fix some element  $s_{\hat{\theta}} \in \hat{S}$ . Because of (73),  $\mathbf{T}(\mathbf{X}, \mathbf{s}_{\hat{\theta}}, \mathbf{s}_{\theta'}) = 0$  for  $s_{\hat{\theta}'} \in \hat{S}$  and, for  $s_{\theta'} \in S \setminus \hat{S}$ ,

$$\mathbf{T}(\mathbf{X}, \mathbf{s}_{\hat{\theta}}, \mathbf{s}_{\theta'}) = \sum_{i=1}^n \frac{\sqrt{s_{\theta'}(X_i)} - \sqrt{s_{\hat{\theta}}(X_i)}}{\sqrt{2(s_{\theta'}(X_i) + s_{\hat{\theta}}(X_i))}} \leq \frac{\sqrt{s_{\theta'}(X_{i_0})} - \sqrt{s_{\hat{\theta}}(X_{i_0})}}{\sqrt{2(s_{\theta'}(X_{i_0}) + s_{\hat{\theta}}(X_{i_0}))}} = -\frac{1}{\sqrt{2}}.$$

Consequently  $\Upsilon(S, \mathbf{s}_{\hat{\theta}}) = 0$  and therefore  $\mathbf{s}_{\hat{\theta}}$  minimizes  $\Upsilon(S, \cdot)$  over  $S$ .  $\square$

We shall prove in Section 9.9, the following result.

**Proposition 30.** *For all  $s \in \bar{S}$  and  $\bar{s} \in S$ ,*

$$D^S(\mathbf{s}, \bar{\mathbf{s}}) \leq 4c_0^{-2} \left[ 2\sqrt{2} + n^2 h^2(s, \bar{s}) \right]^2.$$

Applying (26) we get, since  $\mathbf{h}^2(\cdot, \cdot) = nh^2(\cdot, \cdot)$ ,

$$\mathbb{E}_{\mathbf{s}} [h^2(s, \hat{s})] \leq c_1 h^2(s, \bar{s}) + \frac{4c_2 c_0^{-2} (2\sqrt{2} + n^2 h^2(s, \bar{s}))^2 + 2.45c_3}{n} \quad \text{for all } s \in \bar{S} \text{ and } \bar{s} \in S.$$

Choosing  $\bar{s}$  arbitrarily close to  $s \in \bar{S}$  shows that any  $\rho$ -estimator  $\hat{s}$  (and therefore any maximum likelihood estimator) satisfies

$$\mathbb{E}_{\mathbf{s}} [h^2(s, \hat{s})] \leq (32c_2 c_0^{-2} + 2.45c_3) n^{-1} \quad \text{for all } s \in \bar{S}$$

and thus achieves a parametric rate of convergence independently of the metric dimension of  $\bar{S}$ .

## 7. MODEL SELECTION

Let us now assume that, in place of a single model as in the previous sections, we have at disposal a countable collection  $\bar{\mathbb{S}}$  of such models  $\bar{S}$  for the parameter  $s$ . We may therefore associate to each  $\bar{S} \in \bar{\mathbb{S}}$  a  $\rho$ -estimator  $\hat{s}(\bar{S})$  with quadratic risk  $\mathbb{E}_{\mathbf{s}} [\mathbf{h}^2(\mathbf{s}, \hat{s}(\bar{S}))]$  and our aim is to select from the data  $\mathbf{X}$  a model  $\hat{S} \in \bar{\mathbb{S}}$  or, equivalently, an estimator  $\hat{s}(\hat{S})$  among the family of candidates  $\{\hat{s}(\bar{S}), \bar{S} \in \bar{\mathbb{S}}\}$ , in such a way that its risk is as close as possible to the minimal risk over the family, namely  $\inf_{\bar{S} \in \bar{\mathbb{S}}} \mathbb{E}_{\mathbf{s}} [\mathbf{h}^2(\mathbf{s}, \hat{s}(\bar{S}))]$ .

**7.1. Estimation procedure and main result.** Let  $\bar{\mathbb{S}}$  be a countable family of models in  $\mathcal{L}_0$ , endowed with a mapping  $\Delta$  from  $\bar{\mathbb{S}}$  into  $\mathbb{R}_+$  satisfying

$$(75) \quad \sum_{\bar{S} \in \bar{\mathbb{S}}} \exp[-\Delta(\bar{S})] \leq 1.$$

To each  $\bar{S} \in \bar{\mathbb{S}}$  we attach some identifiable subset  $S$  of  $\mathcal{L}_0$  which is either a countable  $\eta$ -net for  $\bar{S}$  or a dense subset of  $\bar{S}$ , as we did for a single model in Section 4.5, the connection between  $\bar{S}$  and  $S$  being only emphasized by the notations. This results in a new collection

$\mathbb{S}$  of subsets  $S$  of  $\mathcal{L}_0$ , each  $S \in \mathbb{S}$  corresponding to a model  $\bar{S} \in \bar{\mathbb{S}}$  and the set  $\mathbf{S} = \bigcup_{\bar{S} \in \bar{\mathbb{S}}} S = \bigcup_{S \in \mathbb{S}} S$  is a countable subset of  $\mathcal{L}_0$ . Let  $\text{pen}$  be some positive function on  $\mathbf{S}$ ,

$$(76) \quad \bar{\mathbf{T}}(\mathbf{S}, \mathbf{t}) = \sup_{\mathbf{t}' \in \mathbf{S}} \{ \mathbf{T}(\mathbf{X}, \mathbf{t}, \mathbf{t}') - \text{pen}(\mathbf{t}') \} + \text{pen}(\mathbf{t}) \quad \text{for all } \mathbf{t} \in \mathbf{S}$$

and

$$\mathcal{E}(\mathbf{X}, \mathbf{S}) = \left\{ \tilde{\mathbf{s}} \in \mathbf{S} \mid \bar{\mathbf{T}}(\mathbf{S}, \tilde{\mathbf{s}}) \leq \inf_{\mathbf{t} \in \mathbf{S}} \bar{\mathbf{T}}(\mathbf{S}, \mathbf{t}) + \frac{\kappa}{10} \right\} \quad \text{with } \kappa \text{ given by (8).}$$

As in Section 4.2, we define our estimator  $\hat{\mathbf{s}}$  of  $\mathbf{s}$  as any element of  $\text{Cl}(\mathcal{E}(\mathbf{X}, \mathbf{S}))$ . When  $\bar{\mathbb{S}}$  reduces to a single element  $\bar{S}$  and the function  $\text{pen}$  is constant on  $S$ , the estimator  $\hat{\mathbf{s}}$  coincides with the one we defined in Section 4.2.

It follows from (76) that the estimator only depends on the differences  $\text{pen}(\mathbf{t}') - \text{pen}(\mathbf{t})$  rather than on the function  $\text{pen}$  itself. This means that, in our computations, we may always replace the actual penalty function  $\text{pen}$  that has been used to build the  $\rho$ -estimator by another one,  $\text{pen}'$ , with  $\text{pen}'(\mathbf{t}) = \text{pen}(\mathbf{t}) + G$  for some  $G$  independent of  $\mathbf{t} \in \mathbf{S}$ .

**Theorem 31.** *If the penalty function  $\text{pen}$  satisfies*

$$(77) \quad \text{pen}(\mathbf{t}) \geq \text{pen}_1(\mathbf{t}) = \inf_{\{S \in \mathbb{S} \mid S \ni \mathbf{t}\}} \left\{ (1/8) \bar{D}^S + \kappa \Delta(\bar{S}) \right\} \quad \text{for all } \mathbf{t} \in \mathbf{S},$$

where  $\bar{D}^S$  is defined by (28), any element  $\hat{\mathbf{s}}$  in  $\text{Cl}(\mathcal{E}(\mathbf{X}, \mathbf{S}))$  satisfies, for all  $\mathbf{s} \in \mathcal{L}_0$  and  $\xi > 0$ ,

$$(78) \quad \mathbb{P}_{\mathbf{s}} \left[ \mathbf{h}^2(\mathbf{s}, \hat{\mathbf{s}}) \leq \inf_{\bar{S} \in \bar{\mathbb{S}}} \{ c_1 \mathbf{h}^2(\mathbf{s}, \bar{S}) + 8c_2 \text{pen}(\bar{S}) \} + c_3(1.45 + \xi) \right] \geq 1 - e^{-\xi}.$$

In particular, if  $\text{pen}(\mathbf{t}) = \text{pen}_1(\mathbf{t})$  for all  $\mathbf{t} \in \mathbf{S}$  and the sets  $S \in \mathbb{S}$  are chosen in order to satisfy

$$2c_1 \sup_{\mathbf{u} \in \bar{S}} \mathbf{h}^2(\mathbf{u}, S) + c_2 \bar{D}^S \leq \bar{D}(\bar{S}) + \frac{c_3}{20},$$

then for all  $\mathbf{s} \in \mathcal{L}_0$  and  $\xi > 0$ ,

$$(79) \quad \mathbb{P}_{\mathbf{s}} \left[ \mathbf{h}^2(\mathbf{s}, \hat{\mathbf{s}}) \leq \inf_{\bar{S} \in \bar{\mathbb{S}}} \{ 2c_1 \mathbf{h}^2(\mathbf{s}, \bar{S}) + \bar{D}(\bar{S}) + c_3 (\Delta(\bar{S}) + 1.5 + \xi) \} \right] \geq 1 - e^{-\xi}.$$

When the models  $\bar{S}$  are VC-subgraph with respective indices  $\bar{V}(\bar{S})$  we have seen in Theorem 12 that

$$\bar{D}^S \leq C' \bar{V}(\bar{S}) [1 + \log_+(n/\bar{V}(\bar{S}))]$$

for all choices of a countable and dense subset  $S$  of  $\bar{S}$ . For such choices of  $S$  and a penalty equal to  $\text{pen}_1$ , we derive from (78) that the estimator  $\hat{\mathbf{s}}$  satisfies, for all  $\xi > 0$ ,

$$(80) \quad \mathbb{P}_{\mathbf{s}} \left[ C \mathbf{h}^2(\mathbf{s}, \hat{\mathbf{s}}) \leq \inf_{\bar{S} \in \bar{\mathbb{S}}} \left\{ \mathbf{h}^2(\mathbf{s}, \bar{S}) + \bar{V}(\bar{S}) [1 + \log_+(n/\bar{V}(\bar{S}))] + \Delta(\bar{S}) + \xi \right\} \right] \geq 1 - e^{-\xi}.$$

As we have seen in Proposition 17, we are not always able to bound the uniform dimension  $\bar{D}(\bar{S})$  of a model  $\bar{S}$  from above but sometimes only its dimension  $D(\bar{S})$ . In this case, model selection is still possible under the following alternative assumption.

**Assumption 32.** *Let  $\check{D}$  be a mapping from  $\bar{\mathbb{S}}$  into  $[1, +\infty)$  such that for all  $\bar{S}, \bar{S}' \in \bar{\mathbb{S}}$ ,*

$$D^{S \cup S'} \leq \check{D}(\bar{S}) + \check{D}(\bar{S}').$$

**Theorem 33.** *Let Assumption 32 hold and the penalty function  $\text{pen}$  satisfy*

$$(81) \quad \text{pen}(\mathbf{t}) \geq \inf_{\{\bar{S} \in \bar{\mathbb{S}} \mid S \ni \mathbf{t}\}} \left\{ (1/8) \check{D}(\bar{S}) + \kappa \Delta(\bar{S}) \right\} \quad \text{for all } \mathbf{t} \in \mathbf{S}.$$

*The estimator  $\hat{\mathbf{s}}$  then satisfies for all  $\mathbf{s} \in \mathcal{L}_0$  and  $\xi > 0$ ,*

$$(82) \quad \mathbb{P}_{\mathbf{s}} \left[ \mathbf{h}^2(\mathbf{s}, \hat{\mathbf{s}}) \leq \inf_{\bar{\mathbf{s}} \in \bar{\mathbf{S}}} \{c_1 \mathbf{h}^2(\mathbf{s}, \bar{\mathbf{s}}) + 16c_2 \text{pen}(\bar{\mathbf{s}}) + c_3(1.45 + \xi)\} \right] \geq 1 - e^{-\xi}.$$

*If, moreover, the sets  $S \in \bar{\mathbb{S}}$  are chosen to satisfy,*

$$(83) \quad 2c_1 \sup_{\mathbf{u} \in \bar{S}} \mathbf{h}^2(\mathbf{u}, S) + c_2 D^S \leq D(\bar{S}) + \frac{c_3}{80a} \quad \text{for all } \bar{S} \in \bar{\mathbb{S}} \text{ and some } a \geq 1/2,$$

*Assumption 32 holds with  $\check{D}(\bar{S}) \leq aD^S$  and equality holds in (81), then for all  $\mathbf{s} \in \mathcal{L}_0$  and  $\xi > 0$ ,*

$$\mathbb{P}_{\mathbf{s}} \left[ \mathbf{h}^2(\mathbf{s}, \hat{\mathbf{s}}) \leq \inf_{\bar{S} \in \bar{\mathbb{S}}} \{2c_1 \mathbf{h}^2(\mathbf{s}, \bar{S}) + 2aD(\bar{S}) + c_3(2\Delta(\bar{S}) + 1.5 + \xi)\} \right] \geq 1 - e^{-\xi}.$$

We shall now turn to examples in the next sections. Throughout these sections, we shall assume that  $\hat{\mathbf{s}}$  is built with a choice of the penalty function equal to  $\text{pen}_1$  as defined in (77). Finally, given a countable set  $T$ , we shall say that  $\pi$  is a positive sub-probability on  $T$ , if  $\pi(t) > 0$  for all  $t$  in  $T$  and  $\sum_{t \in T} \pi(t) \leq 1$ . Given such a  $\pi$ , we shall set  $\Delta_{\pi}(t) = -\log(\pi(t))$ . It follows that choosing a function  $\Delta$  which satisfies (75) amounts to finding a subprobability  $\pi$  on  $\bar{\mathbb{S}}$  and setting  $\Delta = \Delta_{\pi}$ .

**7.2. Homoscedastic regression with unknown scaling.** We consider here the regression setting described in Section 2.4 where  $\mu$  is the Lebesgue measure and both  $p$  and  $\mathbf{f}$  are unknown. Throughout this section, we shall consider a family  $\mathbb{F}$  of subsets  $F \subset \mathbb{R}^n$  to approximate  $\mathbf{f}$  and a family  $\mathcal{Q}$  of densities  $q$  together with a scaling parameter  $\lambda > 0$  to approximate  $p$  by densities of the form  $q_{0,\lambda}$ . We recall from (9) that  $q_{0,\lambda}(x) = \lambda^{-1}q(x/\lambda)$  for  $\lambda > 0$  and  $x \in \mathbb{R}$  and make the following assumptions.

**Assumption 34.** *The family  $\mathbb{F}$  is a countable family of VC-subgraph classes  $F$  with respective VC-indices  $\bar{V}(F)$  and  $\mathbb{F}$  is endowed with a positive sub-probability  $\pi$ .*

**Assumption 35.** *The family  $\mathcal{Q}$  is countable and endowed with a positive sub-probability  $\gamma$ . For each  $q \in \mathcal{Q}$ , Assumption 24 is satisfied and there exists a non-decreasing function  $\mathbf{w}_q$  from  $[1, 2]$  into  $\mathbb{R}_+$  such that  $\mathbf{w}_q(1) = 0$  and*

$$h^2(q, q_{0,\lambda}) \leq \mathbf{w}_q(\lambda) \quad \text{for all } \lambda \in [1, 2].$$

Given a density  $q$  on  $\mathbb{R}$ , a vector  $\mathbf{g} \in \mathbb{R}^n$  and  $\lambda > 0$ , we define the density  $\mathbf{q}_{\mathbf{g},\lambda}$  with respect to the Lebesgue measure on  $\mathbb{R}^n$  according to (10), that is

$$\mathbf{q}_{\mathbf{g},\lambda}(x_1, \dots, x_n) = (q_{0,\lambda}(x_1 - g_1), \dots, q_{0,\lambda}(x_n - g_n)).$$

We consider the family of models  $\bar{\mathbb{S}}$  defined as follows. For  $i \in \mathbb{N}$ ,  $j \in \mathbb{Z}$ ,  $k \in \{0, \dots, 2^i - 1\}$ ,  $F \in \mathbb{F}$  and  $q \in \mathcal{Q}$ , let

$$\bar{S}_{q,F}^{i,j,k} = \{\mathbf{q}_{\mathbf{g},\lambda} \mid \mathbf{g} \in F, \lambda = \lambda_{i,j,k}\} \quad \text{with} \quad \lambda_{i,j,k} = 2^j(1 + k2^{-i})$$

and define the family  $\bar{\mathbb{S}}$  as

$$(84) \quad \bar{\mathbb{S}} = \left\{ \bar{S}_{q,F}^{i,j,k} \mid q \in \mathcal{Q}, F \in \mathbb{F}, (i, j, k) \in \mathbb{N} \times \mathbb{Z} \times \{0, \dots, 2^i - 1\} \right\}.$$

We endow  $\overline{\mathbb{S}}$  with the weights  $\Delta$  given by

$$(85) \quad \Delta\left(\overline{S}_{q,F}^{i,j,k}\right) = \Delta_\gamma(q) + \Delta_\pi(F) + |j| + i + 2 + i \log 2$$

and we check that

$$\sum_{\overline{S} \in \overline{\mathbb{S}}} e^{-\Delta(\overline{S})} = \sum_{q \in \mathcal{Q}} \gamma(q) \sum_{F \in \mathbb{F}} \pi(F) \sum_{j \in \mathbb{Z}} e^{-(|j|+1)} \sum_{i \in \mathbb{N}} e^{-(i+1)} \sum_{k=0}^{2^i-1} e^{-i \log 2} < 1.$$

Then, the following holds.

**Theorem 36.** *Let  $\mathbb{F}$  be a family of models satisfying Assumption 34 and  $\mathcal{Q}$  a family of densities satisfying Assumption 35. For the collection  $\overline{\mathbb{S}}$  and weight function  $\Delta$  defined by (84) and (85) respectively, any  $\rho$ -estimator  $\widehat{\mathbf{s}}$  satisfies for all densities  $p$  and parameters  $\mathbf{f} \in \mathbb{R}^n$ ,*

$$\begin{aligned} C\mathbb{E}_{\mathbf{s}} [\mathbf{h}^2(\mathbf{s}, \widehat{\mathbf{s}})] &\leq \inf_{q \in \mathcal{Q}, \lambda > 0} \left[ nh^2(p, q_{0,\lambda}) + \inf_{i \geq 0} [n\mathbf{w}_q(1 + 2^{-i}) + i] + \Delta_\gamma(q) + |\log \lambda| \right] \\ &\quad + \inf_{F \in \mathbb{F}} \left[ \inf_{\mathbf{g} \in F} \mathbf{h}^2(\mathbf{p}_{\mathbf{f}}, \mathbf{p}_{\mathbf{g}}) + \overline{V}(F) [1 + \log_+(n/\overline{V}(F))] + \Delta_\pi(F) \right]. \end{aligned}$$

*Proof.* Applying Theorem 31 to the family  $\overline{\mathbb{S}}$  and using the fact that  $\overline{D}$  is bounded from below by  $c_2$ , we derive that whatever the choices of  $F \in \mathbb{F}$ ,  $\mathbf{g} \in F$ ,  $i \in \mathbb{N}$ ,  $j \in \mathbb{Z}$  and  $k \in \{0, \dots, 2^i - 1\}$ ,  $\mathbb{E}_{\mathbf{s}} [\mathbf{h}^2(\mathbf{s}, \widehat{\mathbf{s}})] = \mathbb{E}_{\mathbf{s}} [\mathbf{h}^2(\mathbf{p}_{\mathbf{f}}, \widehat{\mathbf{s}})] \leq CA(q, F, \mathbf{g}, i, j, k)$  where  $C$  is a universal constant and

$$A(q, F, \mathbf{g}, i, j, k) = \mathbf{h}^2(\mathbf{p}_{\mathbf{f}}, \mathbf{q}_{\mathbf{g}, \lambda_{i,j,k}}) + \overline{D}\left(\overline{S}_{q,F}^{i,j,k}\right) + \Delta_\gamma(q) + \Delta_\pi(F) + |j| + i.$$

By the triangular inequality, for  $\lambda > 0$  and  $\lambda' = \lambda_{i,j,k}$ ,

$$\mathbf{h}(\mathbf{p}_{\mathbf{f}}, \mathbf{q}_{\mathbf{g}, \lambda'}) \leq \mathbf{h}(\mathbf{p}_{\mathbf{f}}, \mathbf{p}_{\mathbf{g}}) + \mathbf{h}(\mathbf{p}_{\mathbf{g}}, \mathbf{q}_{\mathbf{g}, \lambda}) + \mathbf{h}(\mathbf{q}_{\mathbf{g}, \lambda}, \mathbf{q}_{\mathbf{g}, \lambda'}).$$

Since, for densities with respect to the Lebesgue measure, the Hellinger distance is translation and scale invariant,

$$\mathbf{h}^2(\mathbf{p}_{\mathbf{g}}, \mathbf{q}_{\mathbf{g}, \lambda}) = nh^2(p, q_{0,\lambda}) \quad \text{and} \quad \mathbf{h}(\mathbf{q}_{\mathbf{g}, \lambda}, \mathbf{q}_{\mathbf{g}, \lambda'}) = \mathbf{h}(\mathbf{q}_{\mathbf{g}, \lambda/\lambda'}, \mathbf{q}_{\mathbf{g}}) = \sqrt{nh}(q_{0,\lambda/\lambda'}, q).$$

To bound  $\mathbf{h}(\mathbf{p}_{\mathbf{f}}, \mathbf{q}_{\mathbf{g}, \lambda_{i,j,k}})$ , it remains to bound  $h(q_{0,\lambda/\lambda'}, q)$  when  $\lambda' = \lambda_{i,j,k}$ . Let  $j \in \mathbb{Z}$  be such that  $2^j \leq \lambda < 2^{j+1}$ . Then  $|j| \leq |\log \lambda|/(\log 2) + 1$  and for all  $i \in \mathbb{N}$  one can find  $k \in \{0, \dots, 2^i - 1\}$  such that

$$\lambda_{i,j,k} = 2^j [1 + k2^{-i}] \leq \lambda < \lambda_{i,j,k+1} = 2^j [1 + (k+1)2^{-i}],$$

hence  $1 \leq \lambda/\lambda' \leq 1 + 2^{-i} \leq 2$ . It then follows from Assumption 35 that

$$h^2(q, q_{0,\lambda/\lambda'}) \leq \mathbf{w}_q(\lambda/\lambda') \leq \mathbf{w}_q(1 + 2^{-i})$$

since  $\mathbf{w}_q$  is non-decreasing. Putting all these bounds together for these choices of  $i, j, k$  we derive that, for some universal constant  $C'$  and all  $i \in \mathbb{N}$ ,

$$\begin{aligned} C'A(q, F, \mathbf{g}, i, j, k) &\leq \mathbf{h}^2(\mathbf{p}_{\mathbf{f}}, \mathbf{p}_{\mathbf{g}}) + nh^2(p, q_{0,\lambda}) + n\mathbf{w}_q(1 + 2^{-i}) \\ &\quad + \overline{D}\left(\overline{S}_{q,F}^{i,j,k}\right) + \Delta_\gamma(q) + \Delta_\pi(F) + \log |\lambda| + i. \end{aligned}$$



It remains to bound  $\overline{D}(\overline{S}_{q,F}^{i,j,k})$ . Since  $q$  satisfies Assumption 24, so does the density  $q_{0,\lambda'}$  and it follows from Theorem 25 that, under Assumptions 34 and 35,

$$\overline{D}(\overline{S}_{q,F}^{i,j,k}) \leq C'' \overline{V}(F) [1 + \log_+(n/\overline{V}(F))],$$

which concludes the proof.  $\square$

Let us now comment on this result. To fix up the ideas, let us take for  $q$  and  $\lambda$  the values that provide the best approximation of  $p$  by  $q_{0,\lambda}$  among all choices in  $\mathcal{Q} \times (0, +\infty)$ , even though this choice might not be the optimal one in view of minimizing our risk bound. The quantity  $nh^2(p, q_{0,\lambda})$  therefore corresponds to the usual bias term resulting from the approximation of  $p$  by the family of densities  $q'_{0,\lambda'}$  for  $q' \in \mathcal{Q}$  and  $\lambda' > 0$ . The quantity

$$\inf_{F \in \mathbb{F}} \left[ \inf_{\mathbf{g} \in F} \mathbf{h}^2(\mathbf{p}_f, \mathbf{p}_g) + \overline{V}(F) [1 + \log_+(n/\overline{V}(F))] + \Delta_\pi(F) \right]$$

is the bound that we would get, if  $p$  were known, for estimating  $\mathbf{f}$  by model selection among the family  $\bigcup_{F \in \mathbb{F}} F$ . Finally, the quantity  $\Delta_\gamma(q) + \inf_{i \geq 0} [n\mathbf{w}_q(1 + 2^{-i}) + i] + |\log \lambda|$  comes from our estimation of  $q$  and  $\lambda$  by model selection.

This regression model includes in particular the case of a known form of the errors corresponding to  $p = \overline{p}_{0,\tau}$  where  $\overline{p}$  is known and  $\tau$  unknown, in which case it is natural to take  $\mathcal{Q} = \{\overline{p}\}$ . The risk bound then becomes after a proper rescaling:

$$\begin{aligned} C\mathbb{E}_s[\mathbf{h}^2(\mathbf{s}, \widehat{\mathbf{s}})] &\leq \inf_{F \in \mathbb{F}} \left[ \inf_{\mathbf{g} \in F} \mathbf{h}^2(\mathbf{p}_f, \mathbf{p}_g) + \overline{V}(F) [1 + \log_+(n/\overline{V}(F))] + \Delta_\pi(F) \right] \\ &\quad + \inf_{\lambda > 0} [nh^2(\overline{p}_{0,\tau}, \overline{p}_{0,\lambda}) + |\log \lambda|] + \inf_{i \geq 0} [n\mathbf{w}_{\overline{p}}(1 + 2^{-i}) + i] \\ &= \inf_{F \in \mathbb{F}} \left[ \inf_{\mathbf{g} \in F} \mathbf{h}^2(\mathbf{p}_f, \mathbf{p}_g) + \overline{V}(F) [1 + \log_+(n/\overline{V}(F))] + \Delta_\pi(F) \right] \\ &\quad + \inf_{\sigma > 0} [nh^2(\overline{p}, \overline{p}_{0,\sigma}) + |\log \tau \sigma|] + \inf_{i \geq 0} [n\mathbf{w}_{\overline{p}}(1 + 2^{-i}) + i]. \end{aligned} \tag{86}$$

**7.3. An example.** Let us consider the family of densities  $\overline{\mathcal{Q}} = \{p^\beta \mid \beta \geq 0\}$  indexed by the parameter  $\beta \in [0, +\infty)$  and given by

$$(87) \quad p^\beta(x) = \Lambda(\beta)e^{-|x|^{1/\beta}}, \quad \Lambda(\beta) = \left[ \int_{\mathbb{R}} e^{-|x|^{1/\beta}} dx \right]^{-1} \quad \text{for } \beta > 0 \quad \text{and} \quad p^0 = \frac{1}{2} \mathbb{1}_{[-1,1]}.$$

It follows from symmetry and a change of variables that

$$(88) \quad [\Lambda(\beta)]^{-1} = 2 \int_0^\infty e^{-x^{1/\beta}} dx = 2\beta\Gamma(\beta) \quad \text{for } \beta > 0.$$

The family  $\overline{\mathcal{Q}}$  contains the Laplace and Gaussian distributions as well as the uniform distribution on  $[-1, 1]$  which corresponds to the limit of the densities  $p^\beta$  when  $\beta$  tends to 0. We recall from Section 6.1 that these densities are of order  $2/\beta$  for  $\beta > 2$  and of order 1 for  $\beta < 2$ . In particular, for such densities inequality (67) is satisfied with  $\overline{\alpha} = (2/\beta) \wedge 1$  in place of  $\alpha$ . We shall consider  $\overline{\mathcal{Q}}$  as a model for our unknown density  $p$ . In order to apply Theorem 36, which only holds for a countable family  $\mathcal{Q}$ , we have to discretize  $\overline{\mathcal{Q}}$ . To do so, we need the following approximation result the proof of which is postponed to Section 9.10.

**Proposition 37.** For all  $\beta > \beta' > 0$ ,

$$(89) \quad h^2(p^\beta, p^{\beta'}) \leq \begin{cases} (13/6)[(\beta/\beta') - 1]^2 & \text{if } 0 < \beta' < \beta \leq 1, \\ (7/4)(\beta - \beta')^2 & \text{if } 1 < \beta' < \beta \leq 3, \\ [1.3(\beta - \beta')(\log \beta)]^2 & \text{if } 3 < \beta' < \beta. \end{cases}$$

Moreover,

$$(90) \quad h^2(p^\beta, p^0) \leq \beta/2 \quad \text{for } 0 < \beta \leq 1$$

and  $\mathbf{w}_{p^\beta}(\lambda) \leq (3/5)(\lambda - 1)$  for all  $\lambda \in [1, 2]$  and  $\beta \geq 0$ .

We are now in a position to prove the following result.

**Corollary 38.** There exists a countable subset  $\mathcal{Q}$  of  $\overline{\mathcal{Q}}$  and a positive sub-probability  $\gamma$  on  $\mathcal{Q}$  with the following properties: for all  $p^\beta \in \overline{\mathcal{Q}}$  there exists  $p^b \in \mathcal{Q}$  such that

$$(91) \quad b \leq \beta, \quad h^2(p^\beta, p^b) \leq n^{-1}$$

and, for all  $p^b \in \mathcal{Q}$  and a suitable positive constant  $c$  (independent of  $n$ ),

$$(92) \quad \gamma(p^b) = \begin{cases} c(\sqrt{n} \log n)^{-1} & \text{if } 0 \leq b \leq 3, \\ cn^{-1}(b-3)^{-2} & \text{if } b > 3. \end{cases}$$

*Proof.* We define  $\mathcal{Q}$  as the image by the application  $\beta \mapsto p^\beta$  of a countable subset  $B = B_1 \cup B_2 \cup B_3$  of  $\mathbb{R}_+$ . We first build  $B_1 = \{b_0 < b_1 < \dots < b_m\} \subset [0, 1]$  with  $b_0 = 0$ ,  $b_1 = 2/n$ ,  $b_{i+1} = b_i \left(1 + \sqrt{6/(13n)}\right)$  for  $1 \leq i \leq m-2$  and  $b_m = 1 \leq b_{m-1} \left(1 + \sqrt{6/(13n)}\right)$ , which defines the value of  $m$ . It follows from (89) and (90) that for any  $\beta \leq 1$ , there exists  $b_i \leq \beta$  with  $h^2(p^\beta, p^{b_i}) \leq n^{-1}$  and, since  $b_{m-1} = 2n^{-1} \left(1 + \sqrt{6/(13n)}\right)^{m-2} < 1$ ,  $m \leq \kappa_1 \sqrt{n} \log n$  for some constant  $\kappa_1$ . We then build  $B_2 = \{b_{m+1} < \dots < b_{m+l-1}\} \subset (1, 3)$  in a similar way with  $b_{i+1} = b_i + 2/\sqrt{7n}$  for  $m \leq i \leq m+l-2$  and  $b_{m+l-1} < 3 \leq b_{m+l-1} + 2/\sqrt{7n}$ . This implies that  $l \leq \kappa_2 \sqrt{n}$  and, for  $\beta \in (1, 3)$ , (91) holds with  $b \in B_2$  by (89). Finally we build  $B_3 \subset [3, +\infty)$  as the infinite sequence  $(b_{m+l+j})_{j \geq 0}$  with

$$b_{m+l+j} = 3 + \frac{j}{1.3\sqrt{n}\alpha_j}, \quad \alpha_j = \log \left(3 + \frac{j}{1.3\sqrt{n}}\right) \quad \text{for } j \geq 0.$$

Since  $\log 3 \leq \alpha_j < \log(3+j)$ , it follows that  $b_{m+l+j}$  goes to infinity with  $j$  and that  $\log(b_{m+l+j}) < \alpha_j$ . Therefore

$$1.3\sqrt{n}(b_{m+l+j+1} - b_{m+l+j}) = \frac{j+1}{\alpha_{j+1}} - \frac{j}{\alpha_j} < \frac{1}{\alpha_{j+1}} < \frac{1}{\log(b_{m+l+j+1})}.$$

It follows from (89) that, for  $\beta \geq 3$ , there exists  $b \in B_3$  such that (91) holds. Since  $|B_1 \cup B_2| = m+l < (\kappa_1 + \kappa_2) \sqrt{n} \log n$  and

$$\frac{1}{n} \sum_{j \geq 1} (b_{m+l+j} - 3)^{-2} = (1.3)^2 \sum_{j \geq 1} \frac{\alpha_j^2}{j^2} < (1.3)^2 \sum_{j \geq 1} \frac{\log^2(3+j)}{j^2} = \kappa_3 < +\infty,$$

we derive that

$$c^{-1} \gamma \left( \{p^b, b \in \mathcal{Q}\} \right) = \sum_{j=0}^{m+l} (\sqrt{n} \log n)^{-1} + \frac{1}{n} \sum_{j \geq 1} (b_{m+l+j} - 3)^{-2} < (1 + \kappa_1 + \kappa_2 + \kappa_3).$$

This implies that  $\gamma$  is a sub-probability for a large enough value of  $c$ .  $\square$

We may now apply Theorem 36 to our example with the family  $\mathcal{Q}$  and the sub-probability  $\gamma$  provided by Corollary 38, which leads to the following result.

**Corollary 39.** *Let  $\overline{\mathcal{Q}}$  be the family of densities defined by (87) and  $\mathbb{F}$  be a family of models satisfying Assumption 34. Let  $\mathcal{Q}$  and  $\gamma$  be given by Corollary 38. For  $\overline{\mathbb{S}}$  and  $\Delta$  defined by (84) and (85) respectively, the estimator  $\widehat{\mathbf{s}}$  satisfies, for all densities  $p$  and vectors  $\mathbf{f} \in \mathbb{R}^n$ ,*

$$\begin{aligned} C\mathbb{E}_{\mathbf{s}}[\mathbf{h}^2(\mathbf{s}, \widehat{\mathbf{s}})] &\leq \inf_{F \in \mathbb{F}} \left[ \inf_{\mathbf{g} \in F} \mathbf{h}^2(\mathbf{p}_{\mathbf{f}}, \mathbf{p}_{\mathbf{g}}) + \overline{V}(F) \left( 1 + \log_+ \left( \frac{n}{\overline{V}(F)} \right) \right) + \Delta_{\pi}(F) \right] \\ &\quad + \inf_{\beta \geq 0, \lambda > 0} \left[ nh^2(p, p_{0,\lambda}^{\beta}) + \log_+(\beta) + |\log \lambda| \right]. \end{aligned}$$

In particular, if  $p = p_{0,\tau}^{\beta}$  for some  $p^{\beta} \in \overline{\mathcal{Q}}$  with  $\beta \neq 2$  and  $\tau > 0$ ,

$$\begin{aligned} C(\beta)\mathbb{E}_{\mathbf{s}}[\mathbf{h}^2(\mathbf{s}, \widehat{\mathbf{s}})] &\leq \inf_{F \in \mathbb{F}} \left[ \inf_{\mathbf{g} \in F} d_{1+[(2/\beta) \wedge 1]}(\mathbf{f}, \mathbf{g}) + \overline{V}(F) \left( 1 + \log_+ \left( \frac{n}{\overline{V}(F)} \right) \right) + \Delta_{\pi}(F) \right] \\ &\quad + \inf_{\sigma > 0} \left[ nh^2(p^{\beta}, p_{0,\sigma}^{\beta}) + |\log \tau \sigma| \right], \end{aligned}$$

where the distance  $d_{1+\alpha}$  has been defined in (68).

*Proof.* Clearly, the family  $\mathcal{Q}$  satisfies Assumption 35, the last requirement deriving from Proposition 37 with  $\mathbf{w}_q(\lambda) = (3/5)(\lambda - 1)$  for all  $\lambda \in [1, 2]$  and  $q \in \mathcal{Q}$ . Under Assumption 34 on the family  $\mathbb{F}$ , we may apply Theorem 36 and we get that, for  $q \in \mathcal{Q}$ ,  $\lambda \in \mathbb{R}_+ \setminus \{0\}$  and  $\mathbf{f} \in \mathbb{R}^n$ , the risk of  $\widehat{\mathbf{s}}$  is bounded by  $C(R_1 + R_2)$  with

$$R_1 = \inf_{F \in \mathbb{F}} \left[ \inf_{\mathbf{g} \in F} \mathbf{h}^2(\mathbf{p}_{\mathbf{f}}, \mathbf{p}_{\mathbf{g}}) + \overline{V}(F) \left( 1 + \log_+ \left( \frac{n}{\overline{V}(F)} \right) \right) + \Delta_{\pi}(F) \right] + \inf_{i \geq 0} [n2^{-i} + i]$$

and

$$R_2 = \inf_{q \in \mathcal{Q}, \lambda > 0} [nh^2(p, q_{0,\lambda}) + \Delta_{\gamma}(q) + |\log \lambda|].$$

Let us first observe that, since  $\overline{V}(F) \geq 1$ ,  $\overline{V}(F) (1 + \log_+ (n/\overline{V}(F))) \geq 1 + \log n$ , so that the term  $\inf_{i \geq 0} [n2^{-i} + i] \leq 2(1 + \log n)$  in  $R_1$  can be ignored at the price of the modification of the universal constant  $C$ . The Hellinger distance being unchanged by scale changes,  $h^2(p, q_{0,\lambda}) \leq 2h^2(p, p_{0,\lambda}^{\beta}) + 2h^2(p^{\beta}, q)$  for any  $p^{\beta} \in \overline{\mathcal{Q}}$  and  $q \in \mathcal{Q}$  so that, with  $q$  chosen in order that  $h^2(p^{\beta}, q) \leq n^{-1}$ ,

$$R_2 \leq \inf_{\beta \geq 0, \lambda > 0} [2nh^2(p, p_{0,\lambda}^{\beta}) + 2 + \Delta_{\gamma}(p^{\beta}) + |\log \lambda|].$$

In view of (91),  $\Delta_{\gamma}(p^{\beta}) \leq C' + \log n + 2\log_+(\beta)$  and the first risk bound follows since we may again omit terms of order  $\log n$ . The second one then derives from the fact that  $p^{\beta}$  is of order  $(2/\beta) \wedge 1$  for  $\beta \neq 2$  and (67), arguing as we did to get (86).  $\square$

**7.4. Random design regression.** We now turn back to the framework of Section 6.3. The same arguments with Theorem 31 replacing Theorem 3 lead to the following generalization of Theorem 28.

**Theorem 40.** *If  $q$  is unimodal and symmetric and  $\mathbb{F}$  is a family of models for  $f$  satisfying Assumption 34, there exists a  $\rho$ -estimator  $\widehat{s} = q_{\widehat{f}}$  of  $s = p_f$  such that for all  $\xi > 0$ , with*

probability at least  $1 - e^{-\xi}$ ,

$$\begin{aligned} Ch^2(p_f, q_{\hat{f}}) &\leq \inf_{F \in \mathbb{F}} \left\{ \inf_{g \in F} h^2(p_f, q_g) + \frac{\overline{V}(F)}{n} \left[ 1 + \log_+ \left( \frac{n}{\overline{V}(F)} \right) \right] + \frac{\Delta_\pi(F)}{n} \right\} + \frac{\xi}{n} \\ &\leq 2h^2(p, q) + \inf_{F \in \mathbb{F}} \left\{ \inf_{g \in F} 2\ell(f, g) + \frac{\overline{V}(F)}{n} \left[ 1 + \log_+ \left( \frac{n}{\overline{V}(F)} \right) \right] + \frac{\Delta_\pi(F)}{n} \right\} + \frac{\xi}{n} \end{aligned}$$

with  $\ell$  given by (71). In particular, if  $p$  is known, unimodal and symmetric and  $q = p$ ,

$$C\ell(f, \hat{f}) \leq \inf_{F \in \mathbb{F}} \left\{ \inf_{g \in F} \ell(f, g) + \frac{\overline{V}(F)}{n} \left[ 1 + \log_+ \left( \frac{n}{\overline{V}(F)} \right) \right] + \frac{\Delta_\pi(F)}{n} \right\} + \frac{\xi}{n}$$

with probability at least  $1 - e^{-\xi}$ .

If, moreover, (67) holds and  $\max \{ \sup_{g \in \mathcal{F}} \|g\|_\infty, \|f\|_\infty \} \leq b < +\infty$ , then

$$C' \|f - \hat{f}\|_{1+\alpha, \nu}^{1+\alpha} \leq \inf_{F \in \mathbb{F}} \left\{ \inf_{g \in F} \|f - g\|_{1+\alpha, \nu}^{1+\alpha} + \frac{\overline{V}(F)}{n} \left[ 1 + \log_+ \left( \frac{n}{\overline{V}(F)} \right) \right] + \frac{\Delta_\pi(F)}{n} \right\} + \frac{\xi}{n}$$

with probability at least  $1 - e^{-\xi}$  for some constant  $C'$  depending only on  $A_p, a_p, b$  and  $\alpha$ .

We are not aware of any other procedure that leads to a comparable result. To illustrate this fact, let us consider the following example. We assume that  $p$  is approximately known (approximately equal to  $q$ ) and that the regression function  $f$  takes the form

$$f = \Psi(\zeta) \quad \text{with} \quad \zeta = \sum_{j=1}^M \beta_j \zeta_j, \quad \beta = (\beta_1, \dots, \beta_M) \in \mathbb{R}^M,$$

where the  $\zeta_j$  are  $M$  given functions on  $\mathcal{W}$ ,  $\Psi$  is an unknown non-decreasing function on  $\mathbb{R}$  and  $M$  may be larger than  $n$  but most of the coefficients  $\beta_j$  are equal to zero, which means that we ignore which functions  $\zeta_j$  are really influential. We also choose a finite set  $\{\Psi_k, 1 \leq k \leq K\}$  of non-decreasing functions to approximate  $\Psi$ .

Given  $k \in \{1, \dots, K\}$  and a non-void subset  $m$  of  $\mathcal{M} = \{1, \dots, M\}$ , we consider the model

$$F_{k,m} = \left\{ \Psi_k \left( \sum_{j \in m} \beta_j \zeta_j \right), \beta_j \in \mathbb{R} \quad \text{for all } j \in \mathcal{M} \right\}.$$

This leads to the family  $\mathbb{F} = \{F_{k,m}, 1 \leq k \leq K \text{ and } m \subset \mathcal{M}\}$  of models for  $f$  and we may set  $\pi(F_{k,m}) = [KM(eM/|m|)^{|m|}]^{-1}$ , so that

$$\sum_{k=1}^K \sum_{m \in \mathcal{M}} \pi(F_{k,m}) = \sum_{k=1}^K \sum_{l=1}^M \sum_{\{m \in \mathcal{M} \mid |m|=l\}} \frac{1}{KM} \left( \frac{eM}{l} \right)^{-l} \leq 1,$$

since  $\binom{M}{l} \leq (eM/l)^l$ . It follows after some simplifications, that the quadratic risk of the corresponding  $\rho$ -estimator can be bounded in the following way (since  $n \geq 3$ ):

$$\begin{aligned} C\mathbb{E}_s \left[ h^2(p_f, q_{\hat{f}}) \right] &\leq h^2(p, q) + \inf_{1 \leq k \leq K} \inf_{m \in \mathcal{M}} \left[ \inf_{g \in F_{k,m}} \ell(\Psi(\zeta), g) + \frac{|m|}{n} \log \left( \frac{nM}{|m|} \right) \right] + \frac{\log(KM)}{n}. \end{aligned}$$

## 8. VC-CLASSES AND SUBGRAPHS

We recall, following Dudley (1984) that

**Definition 41.** Let  $\mathcal{C}$  be a non-empty class of subsets of a set  $\Xi$ . If  $A \subset \Xi$  with  $|A| = n$ , then

$$\Delta_n(\mathcal{C}, A) = |\{A \cap B, B \in \mathcal{C}\}| \quad \text{and} \quad \Delta_n(\mathcal{C}) = \max_{A \subset \Xi, |A|=n} \Delta_n(\mathcal{C}, A).$$

If  $V = \sup \{n \in \mathbb{N} \mid \Delta_n(\mathcal{C}) = 2^n\} < +\infty$ , then  $\mathcal{C}$  is a VC-class with VC-dimension  $V$  and VC-index  $\bar{V} = \inf \{n \in \mathbb{N} \mid \Delta_n(\mathcal{C}) < 2^n\} = V + 1$ .

A class  $\mathcal{F}$  of functions from a set  $\mathcal{X}$  with values in  $(-\infty, +\infty]$  is VC-subgraph with dimension  $V$  and index  $\bar{V}$  if the class of subgraphs  $\{(x, u) \in \mathcal{X} \times \mathbb{R}, f(x) > u\}$  as  $f$  varies among  $\mathcal{F}$  is a VC-class of sets in  $\mathcal{X} \times \mathbb{R}$  with dimension  $V$  and index  $\bar{V}$ .

It immediately follows from this definition that any subset of a VC-subgraph class with index  $\bar{V}$  is VC-subgraph with index not larger than  $\bar{V}$ , a property that we shall repeatedly use. Other known properties of VC-subgraph classes directly derive from the properties of VC-classes as described in van der Vaart and Wellner (1996), Lemma 2.6.17.

If  $\mathcal{F}$  is VC-subgraph with index  $\bar{V}$  on a set  $\mathcal{X}$  and  $\|f\|_\infty \leq 1$  for all  $f \in \mathcal{F}$ , it follows from Theorem 2.6.7 in van der Vaart and Wellner (1996) that, for some numerical constant  $K$  and all probability measures  $Q$  on  $\mathcal{X}$ ,

$$(93) \quad N(\mathcal{F}, Q, \epsilon) \leq K \bar{V} (16e)^{\bar{V}+1} \epsilon^{-2(\bar{V}-1)} \quad \text{for } 0 < \epsilon < 1.$$

Noticing that  $N(\mathcal{F}, Q, 1) = 1$  since the closed ball of center 0 and radius 1 contains  $\mathcal{F}$ , we derive, since  $\bar{V} \geq 1$ , that there exists a universal constant  $A$  such that

$$(94) \quad \log N(\mathcal{F}, Q, \epsilon) \leq 2\bar{V} \log_+(A/\epsilon) \quad \text{for all } \epsilon > 0.$$

When the functions lying in  $\mathcal{F}$  are all non-negative, it is not difficult to see that we can restrict the class of subgraphs to that of “non-negative subgraphs” gathering the sets of the form  $\{(x, u) \in \mathcal{X} \times \mathbb{R}, f(x) > u \geq 0\}$ . If  $S$  is a subset of a linear space of dimension  $D$  then  $S$  is VC-subgraph with index  $\bar{V} \leq D + 2$  (see Lemma 2.6.15 in van der Vaart and Wellner (1996)). We shall repeatedly use the following properties of VC-subgraph classes.

**Proposition 42.** Let  $\mathcal{F}$  be VC-subgraph with dimension  $V$  on a set  $\mathcal{X}$ .

- (i) For all functions  $g$  on  $\mathcal{X}$ ,  $\mathcal{F} + g = \{f + g, f \in \mathcal{F}\}$  is VC-subgraph with dimension not larger than  $V$ .
- (ii) For all monotone function  $\varphi$  on  $\mathbb{R}$ ,  $\varphi(\mathcal{F}) = \{\varphi \circ f, f \in \mathcal{F}\}$  is VC-subgraph with dimension not larger than  $V$ .
- (iii) The class  $-\mathcal{F}$  is VC-subgraph with dimension not larger than  $V$ .
- (iv) The class  $\mathcal{F}_+ = \{f \vee 0, f \in \mathcal{F}\}$  is VC-subgraph with dimension not larger than  $V$ .
- (v) If  $\mathcal{F}$  and  $\mathcal{G}$  are VC-subgraph with respective dimensions  $V$  and  $V'$ ,  $\mathcal{F} \vee \mathcal{G} = \{f \vee g, f \in \mathcal{F}, g \in \mathcal{G}\}$  is VC-subgraph with dimension not larger than  $4.701(V + V')$  and the same holds for  $\mathcal{F} \wedge \mathcal{G} = \{f \wedge g, f \in \mathcal{F}, g \in \mathcal{G}\}$ .
- (vi) If  $q$  is unimodal, the class  $q(\mathcal{F}) = \{q \circ f, f \in \mathcal{F}\}$  is VC-subgraph with dimension not larger than  $9.41V$ .
- (vii) Let  $\psi$  be given by (18),  $g$  be some non-negative function on  $\mathcal{X}$  and all functions in  $\mathcal{F}$  be non-negative. The class of functions  $\psi\left(\sqrt{\mathcal{F}/g}\right) = \left\{\psi(\sqrt{f/g}) \mid f \in \mathcal{F}\right\}$  is VC-subgraph with dimension not larger than  $V$ .

*Proof.* For a proof of (i)–(iv), we refer to Lemma 2.6.18 in van der Vaart and Wellner (1996) and for (v) to the bound (1.2) from van der Vaart and Wellner (2009) together with the relationship between VC-classes and VC-subgraph classes as explained in van der Vaart and Wellner (1996), Section 2.6.5.

For (vi) we argue as follows :  $q$  can be written as  $\varphi_1 \wedge \varphi_2$  where  $\varphi_1$  is non-decreasing and  $\varphi_2$  non-increasing so that  $q \circ f = (\varphi_1 \circ f) \wedge (\varphi_2 \circ f)$ . It follows that  $q(\mathcal{F}) \subset \varphi_1(\mathcal{F}) \wedge \varphi_2(\mathcal{F})$ . The bound then follows from (v).

Let us finally prove (vii). It will be useful here and later on to introduce the function  $\phi$  from  $[0, +\infty]$  to  $[-1, 1]$  given by

$$(95) \quad \phi(x) = \psi(\sqrt{x}), \quad \phi(0/0) = \phi(1) = 0 \quad \text{and} \quad \phi(x/0) = \phi(+\infty) = 1 \quad \text{for all } x > 0,$$

according to the conventions of Section (2.1). Note that  $\phi$  is continuous and increasing, hence one-to-one.

Let  $(x_1, u_1), \dots, (x_m, u_m)$  be  $m \geq 1$  points in  $\mathcal{X} \times (-\infty, +\infty]$  shattered by the subgraphs of  $\phi(\mathcal{F}/g) = \psi(\sqrt{\mathcal{F}/g})$ . It suffices to prove that  $m \leq V$ . First note that we necessarily have  $u_i < 1$  for all  $i$  since  $\phi$  is bounded by 1. In particular for all  $i$ ,  $\phi^{-1}(u_i) < +\infty$ . Besides, because of our convention, we also have  $u_i \geq 0$  for those  $i$  such that  $g(x_i) = 0$ , since otherwise there would be no  $f$  in  $\mathcal{F}$  such that  $\phi(f/g)(x_i) \leq u_i < 0$ . For all  $I \subset \{1, \dots, m\}$  there exists an element  $f$  of  $\mathcal{F}$ , depending on  $I$ , such that  $i$  belongs to  $I$  if and only if  $\phi(f/g)(x_i) > u_i$ . This is equivalent to  $f(x_i) > g(x_i)\phi^{-1}(u_i)$  if  $g(x_i) > 0$  and equivalent to  $f(x_i) > 0$  when  $g(x_i) = 0$  since  $u_i \geq 0$ . In both cases, this is equivalent to  $f(x_i) > g(x_i)\phi^{-1}(u_i)$ . This means that the subgraphs of  $\mathcal{F}$  shatter the set  $\{(x_i, g(x_i)\phi^{-1}(u_i)), i = 1, \dots, m\}$ , which is possible only when  $m \leq V$ .  $\square$

**Remark:** The proof of (vi) extends recursively to multimodal functions with a given number  $k$  of modes by noticing that a function with  $k$  modes can be seen as the supremum of a unimodal function and a multimodal one with  $k - 1$  modes. It follows that if  $q$  is multimodal with  $k$  modes,  $q(\mathcal{F})$  is VC-subgraph with dimension not larger than  $C(k)V$ .

## 9. PROOFS

**9.1. Proofs of Theorem 3, 31 and 33.** All three theorems actually follow from the following (slightly) stronger result.

**Theorem 43.** Let  $\mathbf{s} \in \mathcal{L}_0$ ,  $\bar{\mathbf{s}} \in \mathbf{S}$ ,  $G(\mathbf{s}, \bar{\mathbf{s}})$  be an arbitrary function of  $\mathbf{s}$  and  $\bar{\mathbf{s}}$  and let the penalty function  $\text{pen}$  satisfy

$$(96) \quad \text{pen}(\mathbf{t}) + G(\mathbf{s}, \bar{\mathbf{s}}) \geq \text{pen}_0(\mathbf{t}, \mathbf{s}, \bar{\mathbf{s}}) = \inf_{\{S \in \mathbf{S} \mid S \ni \mathbf{t}\}} \left\{ (1/8)D^S(\mathbf{s}, \bar{\mathbf{s}}) + \kappa \Delta(\bar{S}) \right\} \quad \text{for all } \mathbf{t} \in \mathbf{S}.$$

Then the estimator  $\hat{\mathbf{s}}$  satisfies, for all  $\xi > 0$ ,

$$(97) \quad \mathbb{P}_{\mathbf{s}} [\mathbf{h}^2(\mathbf{s}, \hat{\mathbf{s}}) \leq c_1 \mathbf{h}^2(\mathbf{s}, \bar{\mathbf{s}}) - \mathbf{h}^2(\mathbf{s}, \mathbf{S}) + 8c_2 [\text{pen}(\bar{\mathbf{s}}) + G(\mathbf{s}, \bar{\mathbf{s}})] + c_3(1.45 + \xi)] \geq 1 - e^{-\xi}.$$

**9.1.1. Proof of Theorem 3.** Taking  $\mathbf{S} = \{\bar{S}\}$  hence  $\mathbf{S} = S$ ,  $\Delta(\bar{S}) = 0$ ,  $\text{pen}(\mathbf{t}) = 0$  and  $G(\mathbf{s}, \bar{\mathbf{s}}) = \text{pen}_0(\mathbf{t}, \mathbf{s}, \bar{\mathbf{s}}) = (1/8)D^S(\mathbf{s}, \bar{\mathbf{s}})$  for all  $\mathbf{t} \in S$  we derive from (97) that

$$\mathbb{P}_{\mathbf{s}} [\mathbf{h}^2(\mathbf{s}, \hat{\mathbf{s}}) \leq c_1 \mathbf{h}^2(\mathbf{s}, \bar{\mathbf{s}}) - \mathbf{h}^2(\mathbf{s}, S) + c_2 D^S(\mathbf{s}, \bar{\mathbf{s}}) + c_3(1.45 + \xi)] \geq 1 - e^{-\xi} \quad \text{for all } \xi > 0.$$

Then (25) follows from the fact that this inequality is true for all choices of  $\bar{\mathbf{s}} \in S$ . As to (26) and (27), they follow by integration (see our remark following Theorem 3).

9.1.2. *Proof of Theorem 31.* Inequality (78) is a straightforward consequence of Theorem 43 since, for all  $(\mathbf{s}, \bar{\mathbf{s}}) \in \mathcal{L}_0 \times \mathcal{L}_0$ ,  $D^S(\mathbf{s}, \bar{\mathbf{s}}) \leq \bar{D}^S$ , therefore (77) implies that (96) holds with  $G(\mathbf{s}, \bar{\mathbf{s}}) = 0$  and then (78) follows from (97).

9.1.3. *Proof of Theorem 33.* Let us fix some  $\bar{\mathbf{s}} \in \mathbf{S}$ . There exists  $\bar{S}' \in \bar{\mathbb{S}}$  such that  $\bar{\mathbf{s}} \in S'$  and, by the definition of  $D^S(\cdot, \cdot)$  and Assumption 32,

$$D^S(\mathbf{s}, \bar{\mathbf{s}}) \leq D^{S \cup S'}(\mathbf{s}, \bar{\mathbf{s}}) \leq D^{S \cup S'} \leq \check{D}(\bar{S}) + \check{D}(\bar{S}') \quad \text{for all } \bar{S} \in \bar{\mathbb{S}}.$$

If we set  $G(\mathbf{s}, \bar{\mathbf{s}}) = (1/8)\check{D}(\bar{S}')$ , the penalty function  $\text{pen}$  therefore satisfies for all  $\mathbf{t} \in \mathbf{S}$

$$\begin{aligned} G(\mathbf{s}, \bar{\mathbf{s}}) + \text{pen}(\mathbf{t}) &= (1/8)\check{D}(\bar{S}') + \text{pen}(\mathbf{t}) \\ &\geq (1/8)\check{D}(\bar{S}') + \inf_{\{\bar{S} \in \bar{\mathbb{S}} \mid S \ni \mathbf{t}\}} \left\{ (1/8)\check{D}(\bar{S}) + \kappa\Delta(\bar{S}) \right\} \\ &\geq \inf_{\{S \in \mathbb{S} \mid S \ni \mathbf{t}\}} \left\{ (1/8)D^S(\mathbf{s}, \bar{\mathbf{s}}) + \kappa\Delta(\bar{S}) \right\} = \text{pen}_0(\mathbf{t}, \mathbf{s}, \bar{\mathbf{s}}) \end{aligned}$$

and (96) holds. It therefore follows from Theorem 43 that the estimator  $\hat{\mathbf{s}}$  satisfies, for all  $\xi > 0$  with probability at least  $1 - e^{-\xi}$ ,

$$\begin{aligned} \mathbf{h}^2(\mathbf{s}, \hat{\mathbf{s}}) &\leq c_1 \mathbf{h}^2(\mathbf{s}, \bar{\mathbf{s}}) + 8c_2 [\text{pen}(\bar{\mathbf{s}}) + G(\mathbf{s}, \bar{\mathbf{s}})] + c_3(1.45 + \xi) \\ &= c_1 \mathbf{h}^2(\mathbf{s}, \bar{\mathbf{s}}) + 8c_2 \text{pen}(\bar{\mathbf{s}}) + c_2 \check{D}(\bar{S}') + c_3(1.45 + \xi). \end{aligned}$$

Since this holds for all  $S' \ni \bar{\mathbf{s}}$  and  $\text{pen}(\bar{\mathbf{s}}) \geq \inf_{\{\bar{S}' \in \bar{\mathbb{S}} \mid S' \ni \bar{\mathbf{s}}\}} \{(1/8)\check{D}(\bar{S}')\}$  by (81),

$$(98) \quad \mathbb{P}_{\mathbf{s}} [\mathbf{h}^2(\mathbf{s}, \hat{\mathbf{s}}) \leq c_1 \mathbf{h}^2(\mathbf{s}, \bar{\mathbf{s}}) + 16c_2 \text{pen}(\bar{\mathbf{s}}) + c_3(1.45 + \xi)] \geq 1 - e^{-\xi} \quad \text{for all } \xi > 0.$$

Then (82) follows from the fact that  $\bar{\mathbf{s}}$  is arbitrary in  $\mathbf{S}$ .

Let us now fix some model  $\bar{S} \in \bar{\mathbb{S}}$ , choose  $\mathbf{s}' \in \bar{S}$  such that  $\mathbf{h}^2(\mathbf{s}, \mathbf{s}') \leq \mathbf{h}^2(\mathbf{s}, \bar{S}) + c_3/(160c_1)$  and  $\bar{\mathbf{s}} \in S$  such that  $\mathbf{h}^2(\mathbf{s}', \bar{\mathbf{s}}) \leq \mathbf{h}^2(\mathbf{s}', S) + c_3/(160c_1)$ . It follows that

$$\mathbf{h}^2(\mathbf{s}, \bar{\mathbf{s}}) \leq 2\mathbf{h}^2(\mathbf{s}, \bar{S}) + 2\mathbf{h}^2(\mathbf{s}', S) + c_3/(40c_1).$$

If equality holds in (81) for all  $\mathbf{t} \in \mathbf{S}$  and  $\check{D}(\bar{S}) \leq aD^S$  with  $2a \geq 1$ , it follows from (98) that, for all  $\xi > 0$  with probability at least  $1 - e^{-\xi}$ ,

$$\begin{aligned} \mathbf{h}^2(\mathbf{s}, \hat{\mathbf{s}}) &\leq 2c_1 \mathbf{h}^2(\mathbf{s}, \bar{S}) + 2c_1 \mathbf{h}^2(\mathbf{s}', S) + (c_3/40) + 16c_2 \text{pen}(\bar{\mathbf{s}}) + c_3(1.45 + \xi) \\ &\leq 2c_1 \mathbf{h}^2(\mathbf{s}, \bar{S}) + 2c_1 \mathbf{h}^2(\mathbf{s}', S) + 2ac_2 D^S + 16c_2 \kappa\Delta(\bar{S}) + c_3(1.475 + \xi) \\ &\leq 2c_1 \mathbf{h}^2(\mathbf{s}, \bar{S}) + 2a(2c_1 \mathbf{h}^2(\mathbf{s}', S) + c_2 D^S) + c_3(2\Delta(\bar{S}) + 1.475 + \xi) \\ &\leq 2c_1 \mathbf{h}^2(\mathbf{s}, \bar{S}) + 2aD(\bar{S}) + c_3(2\Delta(\bar{S}) + 1.5 + \xi), \end{aligned}$$

where the last inequality derives from (83). The conclusion follows since  $\bar{S}$  is arbitrary in  $\bar{\mathbb{S}}$ .

9.2. **Proof of Theorem 43.** The proof will be divided into 2 steps.

**Step 1.** Here we prove the following fundamental lemma.

**Lemma 44.** *If the function  $\text{pen}$  satisfies (96), then for all  $\bar{\mathbf{s}} \in \mathcal{L}_0$  and  $\xi > 0$ ,*

$$\begin{aligned} \mathbb{P}_{\mathbf{s}} \left[ \frac{1}{\sqrt{2}} \mathbf{Z}(\mathbf{X}, \bar{\mathbf{s}}, \mathbf{t}) \leq 4c_0 (\mathbf{h}^2(\mathbf{s}, \mathbf{t}) + \mathbf{h}^2(\mathbf{s}, \bar{\mathbf{s}})) + \text{pen}(\mathbf{t}) + G(\mathbf{s}, \bar{\mathbf{s}}) + \kappa(1.4 + \xi) \text{ for all } \mathbf{t} \in \mathbf{S} \right] \\ \geq 1 - e^{-\xi}. \end{aligned}$$



*Proof.* The proof relies on two propositions. The first one presents a version of Talagrand's result on the suprema of empirical processes that is proved in Massart (2007). An alternative solution would be to use Theorem 1.1 of Klein and Rio (2005) instead of (100) below. This would lead to an analogue of (99) with different values of the coefficients of  $v^2$  and  $x$  but not uniformly better.

**Proposition 45.** *Let  $T$  be some finite set,  $U_1, \dots, U_n$  be independent centered random vectors with values in  $\mathbb{R}^T$  and  $Z = \sup_{t \in T} |\sum_{i=1}^n U_{i,t}|$ . If for some positive numbers  $b$  and  $v$ ,*

$$\max_{i=1, \dots, n} |U_{i,t}| \leq b \quad \text{and} \quad \sum_{i=1}^n \mathbb{E}_{\mathbf{s}} [U_{i,t}^2] \leq v^2 \quad \text{for all } t \in T,$$

*then, for all positive  $c$  and  $x$ ,*

$$(99) \quad \mathbb{P} [Z \leq (1+c)\mathbb{E}(Z) + (8b)^{-1}cv^2 + 2(1+8c^{-1})bx] \geq 1 - e^{-x}.$$

*Proof.* The second displayed formula on page 170 of Massart (2007) tells us that

$$(100) \quad \mathbb{P} [Z \leq \mathbb{E}(Z) + 2\sqrt{[2v^2 + 16b\mathbb{E}(Z)]x} + 2bx] \geq 1 - e^{-x}.$$

To derive (99) we use the fact that  $2\sqrt{Ax} \leq A/(16bc^{-1}) + (16bc^{-1})x$  which results in

$$2\sqrt{[2v^2 + 16b\mathbb{E}(Z)]x} \leq 2c(16b)^{-1}v^2 + c\mathbb{E}(Z) + 16bc^{-1}x.$$

□

Even though the result is stated for finite  $T$ , it can easily be extended to countable sets  $T$  by monotone convergence.

The second proposition we need is proved in Baraud (2011) (more precisely, we refer to the proof of his Proposition 3 on page 386 with the difference that in this paper his function  $\psi$  is equal to our function  $\psi$  divided by  $\sqrt{2}$  which involves an additional factor 2 for the control of the function  $\psi^2$  as defined by (18)).

**Proposition 46.** *Let  $\mathbf{X} = (X_1, \dots, X_n)$  be a vector of independent random variables and  $\mathbf{t}, \bar{\mathbf{s}} \in \mathcal{L}_0$ . Then*

$$\mathbb{E}_{\mathbf{s}} \left[ \psi^2 \left( \sqrt{\frac{\mathbf{t}}{\bar{\mathbf{s}}}}(\mathbf{X}) \right) \right] = \sum_{i=1}^n \mathbb{E}_{\mathbf{s}} \left[ \psi^2 \left( \sqrt{\frac{t_i}{\bar{s}_i}}(X_i) \right) \right] \leq 6 [\mathbf{h}^2(\mathbf{s}, \mathbf{t}) + \mathbf{h}^2(\mathbf{s}, \bar{\mathbf{s}})].$$

Let us now turn to the proof of Lemma 44. We fix  $\xi > 0$ ,  $S$  in  $\mathbb{S}$ ,  $\tau = a/(32c_0^2) > 0$  for some positive number  $a$  to be chosen later and we set for all  $j \in \mathbb{N}$ ,

$$y_j^2 = \left(\frac{5}{4}\right)^j [D^S(\mathbf{s}, \bar{\mathbf{s}}) + \tau(\Delta(\bar{S}) + \xi + 1.4)], \quad x_j = \frac{y_j^2}{\tau} \geq \Delta(\bar{S}) + \xi + 1.4 \left(\frac{5}{4}\right)^j,$$

$$B_j^S(\mathbf{s}, \bar{\mathbf{s}}) = \{\mathbf{t} \in S \text{ such that } y_j^2 < \mathbf{h}^2(\mathbf{s}, \mathbf{t}) + \mathbf{h}^2(\mathbf{s}, \bar{\mathbf{s}}) \leq y_{j+1}^2\}$$

and

$$Z_j^S(\mathbf{X}, \bar{\mathbf{s}}) = \frac{1}{\sqrt{2}} \sup_{\mathbf{t} \in B_j^S(\mathbf{s}, \bar{\mathbf{s}})} |\mathbf{Z}(\mathbf{X}, \bar{\mathbf{s}}, \mathbf{t})|.$$

For each  $j \geq 0$ , we may apply Proposition 45 to the supremum  $Z_j^S(\mathbf{X}, \bar{\mathbf{s}})$  by taking  $T = B_j^S(\mathbf{s}, \bar{\mathbf{s}})$  (which is countable as a subset of  $S$ ) and

$$(101) \quad U_{i,\mathbf{t}} = \frac{1}{\sqrt{2}} \left\{ \psi \left( \sqrt{\frac{t_i}{\bar{s}_i}}(X_i) \right) - \mathbb{E}_{\mathbf{s}} \left[ \psi \left( \sqrt{\frac{t_i}{\bar{s}_i}}(X_i) \right) \right] \right\} \quad \text{for all } i = 1, \dots, n.$$

For such a choice, the assumptions of the Proposition 45 are met with  $b = \sqrt{2}$  (since  $\psi$  is bounded by 1) and  $v^2 = 3y_{j+1}^2$  (by Proposition 46 and the definition of  $B_j^S(\mathbf{s}, \bar{\mathbf{s}})$ ). It therefore follows from (99) that, with probability at least  $1 - e^{-x_j}$  and for all  $\mathbf{t} \in B_j^S(\mathbf{s}, \bar{\mathbf{s}})$ ,

$$(102) \quad \frac{1}{\sqrt{2}} \mathbf{Z}(\mathbf{X}, \bar{\mathbf{s}}, \mathbf{t}) \leq Z_j^S(\mathbf{X}, \bar{\mathbf{s}}) \leq (1+c) \mathbb{E}_{\mathbf{s}} [Z_j^S(\mathbf{X}, \bar{\mathbf{s}})] + \frac{3}{8\sqrt{2}} c y_{j+1}^2 + 2\sqrt{2} (1+8c^{-1}) x_j.$$

Since  $B_j^S(\mathbf{s}, \bar{\mathbf{s}}) \subset \mathcal{B}^S(\mathbf{s}, \bar{\mathbf{s}}, y_{j+1})$ , it follows from the definition of  $D^S(\mathbf{s}, \bar{\mathbf{s}})$  and the fact that  $y_{j+1}^2 > D^S(\mathbf{s}, \bar{\mathbf{s}})$  that,

$$\mathbb{E} [Z_j^S(\mathbf{X}, \bar{\mathbf{s}})] \leq 2^{-1/2} \mathbf{w}^S(\mathbf{s}, \bar{\mathbf{s}}, y_{j+1}) \leq 2^{-1/2} c_0 y_{j+1}^2,$$

and, since  $x_j = 4y_{j+1}^2/(5\tau)$ , (102) becomes

$$\frac{1}{\sqrt{2}} \mathbf{Z}(\mathbf{X}, \bar{\mathbf{s}}, \mathbf{t}) \leq \frac{y_{j+1}^2}{\sqrt{2}} \left[ c_0(1+c) + \frac{3c}{8} + \frac{16(1+8c^{-1})}{5\tau} \right].$$

Setting  $c = 16(2\tau)^{-1/2}$ , we get with probability at least  $1 - e^{-x_j}$  and for all  $\mathbf{t} \in B_j^S(\mathbf{s}, \bar{\mathbf{s}})$ ,

$$\begin{aligned} & \frac{1}{\sqrt{2}} \mathbf{Z}(\mathbf{X}, \bar{\mathbf{s}}, \mathbf{t}) - 4c_0 [\mathbf{h}^2(\mathbf{s}, \mathbf{t}) + \mathbf{h}^2(\mathbf{s}, \bar{\mathbf{s}})] \\ & \leq \frac{y_{j+1}^2}{\sqrt{2}} \left[ c_0 \left( 1 + \frac{16}{\sqrt{2\tau}} \right) + \frac{6}{\sqrt{2\tau}} + \frac{16}{5\tau} + \frac{16}{5\sqrt{2\tau}} - \frac{16c_0\sqrt{2}}{5} \right]. \end{aligned}$$

The bracketed factor writes

$$\begin{aligned} & c_0 + \frac{64c_0^2}{\sqrt{a}} + \frac{24c_0}{\sqrt{a}} + \frac{512c_0^2}{5a} + \frac{64c_0}{5\sqrt{a}} - \frac{16c_0\sqrt{2}}{5} \\ & = \frac{c_0}{5} \left[ 5 - 16\sqrt{2} + \frac{184}{\sqrt{a}} + 4(2 - \sqrt{2}) \left( \frac{5}{\sqrt{a}} + \frac{8}{a} \right) \right], \end{aligned}$$

which is negative for  $a = 125.4$ . With this choice of  $a$ ,  $\tau = 62.7(4c_0)^{-2}$  and for all  $j \in \mathbb{N}$ ,

$$\mathbb{P}_{\mathbf{s}} \left[ \frac{1}{\sqrt{2}} \mathbf{Z}(\mathbf{X}, \bar{\mathbf{s}}, \mathbf{t}) - 4c_0 [\mathbf{h}^2(\mathbf{s}, \mathbf{t}) + \mathbf{h}^2(\mathbf{s}, \bar{\mathbf{s}})] < 0 \text{ for all } \mathbf{t} \in B_j^S(\mathbf{s}, \bar{\mathbf{s}}) \right] \geq 1 - e^{-x_j}.$$

Let us now define

$$Z^S(\mathbf{X}, \bar{\mathbf{s}}) = \frac{1}{\sqrt{2}} \sup_{\mathbf{t} \in \mathcal{B}^S(\mathbf{s}, \bar{\mathbf{s}}, y_0)} |\mathbf{Z}(\mathbf{X}, \bar{\mathbf{s}}, \mathbf{t})|$$

and apply Proposition 45 in a similar way to  $Z^S(\mathbf{X}, \bar{\mathbf{s}})$  with  $x = x_0 = y_0^2/\tau$  and  $c = 16\sqrt{3/(10\tau)} = 64c_0/\sqrt{209}$ . We then deduce analogously that, with probability at least

$1 - e^{-x_0}$  and for all  $\mathbf{t} \in \mathcal{B}^S(\mathbf{s}, \bar{\mathbf{s}}, y_0)$ ,

$$\begin{aligned} \frac{1}{\sqrt{2}} \mathbf{Z}(\mathbf{X}, \bar{\mathbf{s}}, \mathbf{t}) &\leq Z^S(\mathbf{X}, \bar{\mathbf{s}}) \leq \frac{y_0^2}{\sqrt{2}} \left[ c_0(1+c) + \frac{3c}{8} + \frac{4(1+8c^{-1})}{\tau} \right] \\ &\leq \frac{4c_0 y_0^2}{\sqrt{2}} \left[ 16c_0 \left( \frac{1}{\sqrt{209}} + \frac{1}{62.7} \right) + \frac{1}{4} + \frac{6}{\sqrt{209}} + \frac{2\sqrt{209}}{62.7} \right] < 0.122 y_0^2. \end{aligned}$$

Since  $\{\mathcal{B}^S(\mathbf{s}, \bar{\mathbf{s}}, y_0), \{B_j^S(\mathbf{s}, \bar{\mathbf{s}}), j \geq 0\}\}$  provides a partition of  $S$ , by putting all these inequalities together we derive that for all  $\mathbf{t} \in S$ ,

$$\frac{1}{\sqrt{2}} \mathbf{Z}(\mathbf{X}, \bar{\mathbf{s}}, \mathbf{t}) - 4c_0 (\mathbf{h}^2(\mathbf{s}, \mathbf{t}) + \mathbf{h}^2(\mathbf{s}, \bar{\mathbf{s}})) < 0.122 y_0^2 < (1/8) D^S(\mathbf{s}, \bar{\mathbf{s}}) + \kappa [\Delta(\bar{S}) + \xi + 1.4],$$

except on a set of probability not larger than

$$e^{-x_0} + \sum_{j \geq 0} e^{-x_j} \leq e^{-\xi - \Delta(\bar{S})} \left[ 2e^{-1.4} + \sum_{j \geq 1} e^{-1.4 \times (5/4)^j} \right] < e^{-\xi - \Delta(\bar{S})}.$$

The result finally extends to all  $\mathbf{t} \in \mathbf{S}$  by summing these bounds over  $S \in \mathbb{S}$  and using (75).  $\square$

**Step 2.** Let us now set, for  $\mathbf{s}, \mathbf{t}, \mathbf{t}' \in \mathcal{L}_0$ ,

$$\mathbf{T}(\mathbf{s}, \mathbf{t}, \mathbf{t}') = \mathbb{E}_{\mathbf{s}} [\mathbf{T}(\mathbf{X}, \mathbf{t}, \mathbf{t}')] = \sum_{i=1}^n T(s_i, t_i, t'_i).$$

Applying inequality (14) to each coordinate  $s_i$ ,  $\bar{s}_i$  and  $t_i$  of  $\mathbf{s}$ ,  $\bar{\mathbf{s}}$  and  $\mathbf{t}$  respectively and summing these inequalities over  $i \in \{1, \dots, n\}$  leads to

$$(103) \quad \mathbf{T}(\mathbf{s}, \bar{\mathbf{s}}, \mathbf{t}) \leq c_2 \mathbf{h}^2(\mathbf{s}, \bar{\mathbf{s}}) - 8c_0 \mathbf{h}^2(\mathbf{s}, \mathbf{t}) \quad \text{for all } \mathbf{s} \in \mathcal{L}_0, \bar{\mathbf{s}} \text{ and } \mathbf{t} \in \mathbf{S}.$$

Let us fix  $\mathbf{s} \in \mathcal{L}_0$  and  $\bar{\mathbf{s}} \in \mathbf{S}$ . Recalling that  $\mathbf{Z}(\mathbf{X}, \bar{\mathbf{s}}, \mathbf{t})/\sqrt{2} = \mathbf{T}(\mathbf{X}, \bar{\mathbf{s}}, \mathbf{t}) - \mathbf{T}(\mathbf{s}, \bar{\mathbf{s}}, \mathbf{t})$ , we deduce from Lemma 44 that, with probability at least  $1 - e^{-\xi}$  and for all  $\mathbf{t} \in \mathbf{S}$ ,

$$(104) \quad \mathbf{T}(\mathbf{X}, \bar{\mathbf{s}}, \mathbf{t}) - \mathbf{T}(\mathbf{s}, \bar{\mathbf{s}}, \mathbf{t}) \leq 4c_0 (\mathbf{h}^2(\mathbf{s}, \mathbf{t}) + \mathbf{h}^2(\mathbf{s}, \bar{\mathbf{s}})) + \text{pen}(\mathbf{t}) + G(\mathbf{s}, \bar{\mathbf{s}}) + \kappa(1.4 + \xi),$$

which, together with (103), leads to

$$(105) \quad \begin{aligned} &\mathbf{T}(\mathbf{X}, \bar{\mathbf{s}}, \mathbf{t}) - \text{pen}(\mathbf{t}) \\ &\leq (4c_0 + c_2) \mathbf{h}^2(\mathbf{s}, \bar{\mathbf{s}}) - 4c_0 \mathbf{h}^2(\mathbf{s}, \mathbf{t}) + G(\mathbf{s}, \bar{\mathbf{s}}) + \kappa(1.4 + \xi) \quad \text{for all } \mathbf{t} \in \mathbf{S}. \end{aligned}$$

Hence, with probability at least  $1 - e^{-\xi}$ ,

$$\begin{aligned} \bar{\mathbf{T}}(\mathbf{S}, \bar{\mathbf{s}}) - \text{pen}(\bar{\mathbf{s}}) &= \sup_{\mathbf{t} \in \mathbf{S}} [\mathbf{T}(\mathbf{X}, \bar{\mathbf{s}}, \mathbf{t}) - \text{pen}(\mathbf{t})] \\ &\leq (4c_0 + c_2) \mathbf{h}^2(\mathbf{s}, \bar{\mathbf{s}}) - 4c_0 \mathbf{h}^2(\mathbf{s}, \mathbf{S}) + G(\mathbf{s}, \bar{\mathbf{s}}) + \kappa(1.4 + \xi) \end{aligned}$$

and it follows from the definitions of  $\hat{\mathbf{s}}$  and  $\bar{\mathbf{T}}(\mathbf{S}, \cdot)$  respectively that

$$(106) \quad \begin{aligned} \bar{\mathbf{T}}(\mathbf{S}, \hat{\mathbf{s}}) &\leq \bar{\mathbf{T}}(\mathbf{S}, \bar{\mathbf{s}}) + (\kappa/10) \\ &\leq (4c_0 + c_2) \mathbf{h}^2(\mathbf{s}, \bar{\mathbf{s}}) - 4c_0 \mathbf{h}^2(\mathbf{s}, \mathbf{S}) + \text{pen}(\bar{\mathbf{s}}) + G(\mathbf{s}, \bar{\mathbf{s}}) + \kappa(1.5 + \xi) \end{aligned}$$

and  $\mathbf{T}(\mathbf{X}, \widehat{\mathbf{s}}, \overline{\mathbf{s}}) + \text{pen}(\widehat{\mathbf{s}}) \leq \overline{\mathbf{T}}(\mathbf{S}, \widehat{\mathbf{s}}) + \text{pen}(\overline{\mathbf{s}})$ . Therefore, using (105) with  $\mathbf{t} = \widehat{\mathbf{s}}$ , the fact that  $\mathbf{T}(\mathbf{X}, \overline{\mathbf{s}}, \widehat{\mathbf{s}}) = -\mathbf{T}(\mathbf{X}, \widehat{\mathbf{s}}, \overline{\mathbf{s}})$  and (106), we derive that, with probability at least  $1 - e^{-\xi}$ ,

$$\begin{aligned} 4c_0 \mathbf{h}^2(\mathbf{s}, \widehat{\mathbf{s}}) &\leq (4c_0 + c_2) \mathbf{h}^2(\mathbf{s}, \overline{\mathbf{s}}) + \mathbf{T}(\mathbf{X}, \widehat{\mathbf{s}}, \overline{\mathbf{s}}) + \text{pen}(\widehat{\mathbf{s}}) + G(\mathbf{s}, \overline{\mathbf{s}}) + \kappa(1.4 + \xi) \\ &\leq (4c_0 + c_2) \mathbf{h}^2(\mathbf{s}, \overline{\mathbf{s}}) + \overline{\mathbf{T}}(\mathbf{S}, \widehat{\mathbf{s}}) + \text{pen}(\overline{\mathbf{s}}) + G(\mathbf{s}, \overline{\mathbf{s}}) + \kappa(1.4 + \xi) \\ &\leq 2(4c_0 + c_2) \mathbf{h}^2(\mathbf{s}, \overline{\mathbf{s}}) - 4c_0 \mathbf{h}^2(\mathbf{s}, \mathbf{S}) + 2 \text{pen}(\overline{\mathbf{s}}) + 2G(\mathbf{s}, \overline{\mathbf{s}}) + 2\kappa(1.45 + \xi), \end{aligned}$$

which leads to the result since  $4c_0 = (4c_2)^{-1}$  and  $c_1 = 2 + c_2/(2c_0)$  by (8).

**9.3. Proof of Proposition 5.** It actually follows from the next one:

**Proposition 47.** *If  $\mathbf{s}$  and  $\overline{\mathbf{s}} \in \mathcal{L}_0$  and  $T(\mathbf{X})$  is such that  $\mathbb{P}_{\overline{\mathbf{s}}}[T(\mathbf{X}) \geq z] \leq ae^{-z}$  for all  $z \geq 0$  and some  $a > 0$ , then*

$$(107) \quad \mathbb{E}_{\mathbf{s}}[T(\mathbf{X})] \leq (1 + \zeta^{-1}) [\log(1 + a\zeta) + \mathbf{K}] \quad \text{for all } \zeta > 0 \quad \text{and} \quad \mathbf{K} = \mathbf{K}(\mathbf{s}, \overline{\mathbf{s}}).$$

*In particular  $\mathbb{E}_{\mathbf{s}}[T(\mathbf{X})] \leq 1 + \zeta_0$  where  $\zeta_0$  is the largest solution of the equation  $\zeta = \log(1 + a\zeta) + \mathbf{K}$  in  $(-a^{-1}, +\infty)$ .*

*Proof.* We start with the following lemma which appears in a slightly different form in Barron (1991).

**Lemma 48.** *Let  $P$  and  $Q$  be two probabilities on  $(\mathcal{X}, \mathcal{A})$  and  $f$  a function from  $(\mathcal{X}, \mathcal{A})$  to  $\mathbb{R}$  such that  $\int_{\mathcal{X}} (f \wedge 0) dP > -\infty$ . Then*

$$\int_{\mathcal{X}} f dP \leq \log \left( \int_{\mathcal{X}} e^f dQ \right) + K(P, Q) \leq \log \left( 1 + \int_0^{+\infty} e^\xi Q[f > \xi] d\xi \right) + K(P, Q).$$

*Proof.* The classical variational formula for the Kullback-Leibler divergence asserts that

$$K(P, Q) = \sup_{g \in \mathcal{G}} \int_{\mathcal{X}} g dP \quad \text{with} \quad \mathcal{G} = \left\{ g : (\mathcal{X}, \mathcal{A}) \rightarrow \mathbb{R} \text{ such that } \int_{\mathcal{X}} e^g dQ = 1 \right\}.$$

For  $g = f - \log \left( \int_{\mathcal{X}} e^f dQ \right)$  which belongs to  $\mathcal{G}$ , we obtain that

$$K(P, Q) \geq \int_{\mathcal{X}} g dP = \int_{\mathcal{X}} f dP - \log \left( \int_{\mathcal{X}} e^f dQ \right)$$

which leads to the first inequality. The second inequality derives from

$$\int_{\mathcal{X}} e^f dQ = \int_0^{+\infty} Q[e^f > t] dt = \int_{-\infty}^{+\infty} Q[f > \xi] e^\xi d\xi \leq 1 + \int_0^{+\infty} e^\xi Q[f > \xi] d\xi.$$

□

To prove Proposition 47 we apply the lemma with  $f = \lambda T$ ,  $0 < \lambda < 1$ ,  $P = \mathbb{P}_{\mathbf{s}}$  and  $Q = \mathbb{P}_{\overline{\mathbf{s}}}$ , getting

$$\lambda \mathbb{E}_{\mathbf{s}}[T(\mathbf{X})] \leq \log \left( 1 + \int_0^{+\infty} e^\xi \mathbb{P}_{\overline{\mathbf{s}}}[\lambda T(\mathbf{X}) > \xi] d\xi \right) + \mathbf{K}.$$

Hence, setting  $\zeta = \lambda/(1 - \lambda) > 0$  so that  $\lambda = \zeta/(\zeta + 1)$ , we get

$$\begin{aligned} \frac{\zeta}{\zeta + 1} \mathbb{E}_{\mathbf{s}}[T(\mathbf{X})] &\leq \log \left( 1 + \int_0^{+\infty} e^\xi \mathbb{P}_{\overline{\mathbf{s}}}[T(\mathbf{X}) > \xi/\lambda] d\xi \right) + \mathbf{K} \\ &\leq \log \left( 1 + a \int_0^{+\infty} \exp[-\xi/\zeta] d\xi \right) + \mathbf{K} = \log(1 + a\zeta) + \mathbf{K}, \end{aligned}$$

which proves (107). The function  $g(\zeta) = \zeta - \log(1+a\zeta) - \mathbf{K}$  is strictly convex on  $(-a^{-1}, +\infty)$  with a minimum equal to  $1 - a^{-1} - \log a - \mathbf{K} \leq 0$  when  $\zeta = 1 - a^{-1}$  and  $g(0) = -\mathbf{K} \leq 0$  so that  $\zeta_0 \geq 0$  (actually  $> 0$  except if  $\mathbf{K} = 0$  and  $a \leq 1$ ) and the bound  $\mathbb{E}_s[T(\mathbf{X})] \leq 1 + \zeta_0$  immediately follows when  $\zeta_0 > 0$ . If  $\zeta_0 = 0$ , then  $\mathbf{K} = 0$  and  $a \leq 1$  so that  $\mathbb{E}_s[T(\mathbf{X})] \leq a$ .  $\square$

To prove Proposition 5 we bound  $\mathbb{E}_s[T(\mathbf{X})] \leq 1 + \zeta_0$  in the following way, setting  $f(x) = x - \log(1+x)$ . We first observe that, since  $\log(1+uv) \leq \log(1+u) + \log(1+v)$  for all  $u, v \geq 0$ ,

$$f(\zeta_0) = \zeta_0 - \log(1 + \zeta_0) \leq c = \log(1+a) + \mathbf{K}$$

and, since  $f$  is increasing on  $[0, +\infty[$ ,  $\zeta_0 \leq f^{-1}(c)$ . Moreover,

$$f\left(c + \log\left(1 + c + \sqrt{2c}\right)\right) - c = \log\left(1 + c + \sqrt{2c}\right) - \log\left(1 + c + \log\left(1 + c + \sqrt{2c}\right)\right)$$

has the same sign as

$$(108) \quad \left(1 + c + \sqrt{2c}\right) - \left(1 + c + \log\left(1 + c + \sqrt{2c}\right)\right) = \sqrt{2c} - \log\left(1 + c + \sqrt{2c}\right),$$

which has the sign of  $\exp[\sqrt{2c}] - (1 + c + \sqrt{2c}) > 0$  since  $c > 0$  and  $e^x > 1 + x + (x^2/2)$  for  $x > 0$ . It follows that  $c < f\left(c + \log\left(1 + c + \sqrt{2c}\right)\right)$  and finally, using again the fact that the right-hand side of (108) is positive,

$$1 + \zeta_0 \leq 1 + f^{-1}(c) < 1 + c + \log\left(1 + c + \sqrt{2c}\right) < 1 + c + \sqrt{2c},$$

which completes the proof of (33). To get (34) we apply (33) to the random variable  $T'(\mathbf{X}) = b[T(\mathbf{X}) - z_0]$ .

**9.4. Proof of Proposition 11.** It follows from the next lemma to be proved afterwards.

**Lemma 49.** *Let  $X_1, \dots, X_n$  be independent random variables defined on a probability space  $(\Omega, \mathcal{A}, \mathbb{P})$  and with values in  $\mathcal{X}$ ,  $\mathcal{F}$  a class of functions on  $\mathcal{X}$  bounded by 1 and  $\mathcal{H}$  a function satisfying Assumption 10. If*

$$(109) \quad \sup_{f \in \mathcal{F}} \sum_{i=1}^n \mathbb{E}[f^2(X_i)] \leq v^2$$

and

$$(110) \quad \log N\left(\mathcal{F}, \frac{1}{n} \sum_{i=1}^n \delta_{X_i(\omega)}, z\right) \leq \overline{\mathcal{H}}\left(\frac{1}{z}\right) \quad \text{for all } \omega \in \Omega \text{ and } 0 < z \leq 2,$$

then there exists a universal constant  $C_0$  such that,

$$\mathbb{E} \left[ \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n (f(X_i) - \mathbb{E}[f(X_i)]) \right| \right] \leq C_0 \left[ vL\sqrt{H} + L^2H \right] \quad \text{with} \quad H = \overline{\mathcal{H}}\left(\frac{\sqrt{n}}{2v} \vee \frac{1}{2}\right).$$

To prove (47) for a given value of  $y$  we use (45) and apply this lemma to the family  $\mathcal{F}^S(\mathbf{s}, \overline{\mathbf{s}}, y)$ , the elements of which are bounded by 1 and satisfy (109) with  $v^2 = 6y^2$  by Proposition 46, and to the function  $\overline{\mathcal{H}}_y$ , so that  $H = H_y$  and  $L = L_y$ . Since  $2ab \leq \alpha a^2 + \alpha^{-1}b^2$  for all  $a, b \in \mathbb{R}$  and  $\alpha = c_0/C_0$ , we derive that

$$\mathbf{w}^S(\mathbf{s}, \overline{\mathbf{s}}, y) \leq \frac{C_0}{2} [\alpha y^2 + (2 + 6\alpha^{-1}) L_y^2 H_y] \leq \frac{c_0 y^2}{2} + C_0 \left(1 + \frac{3C_0}{c_0}\right) L_y^2 H_y.$$

Then (48) follows from the definition of  $D^S(\mathbf{s}, \bar{\mathbf{s}})$ .

Let us now turn to the proof of Lemma 49. The line of proof is the same as that of Theorem 3.1 of Giné and Koltchinskii (2006) with minor changes due to the fact that we consider non i.i.d. random variables  $X_i$ . Similar arguments were used in Massart and Nédélec (2006) for classes  $\mathcal{F}$  of indicator functions.

By a symmetrization argument,

$$\mathbb{E} \left[ \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n (f(X_i) - \mathbb{E}[f(X_i)]) \right| \right] \leq 2E = 2\mathbb{E} \left[ \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \varepsilon_i f(X_i) \right| \right],$$

where the  $\varepsilon_i$  are Rademacher random variables independent of the  $X_i$ . Arguing as in Giné and Koltchinskii with  $F = 1$ , we get

$$E \leq Cn^{1/2} \mathbb{E} \left[ \int_0^{2\hat{\sigma}_n} \sqrt{\mathcal{H}(1/z)} dz \right] \quad \text{with} \quad \hat{\sigma}_n^2 = \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n f^2(X_i) \leq 1.$$

The function  $u \mapsto \mathcal{H}(1/u)$  being non-increasing,  $u \mapsto \int_0^u \mathcal{H}(1/z) dz$  is concave and therefore

$$E \leq Cn^{1/2} \int_0^{2\mathbb{E}[\hat{\sigma}_n]} \sqrt{\mathcal{H}(1/z)} dz \leq Cn^{1/2} \int_0^{2\sqrt{\mathbb{E}[\hat{\sigma}_n^2]}} \sqrt{\mathcal{H}(1/z)} dz.$$

Symmetrization and contraction arguments together with the fact that  $|f| \leq 1$  for all  $f \in \mathcal{F}$  lead to

$$\mathbb{E}[\hat{\sigma}_n^2] \leq B^2 = \frac{v^2 + 8E}{n} \wedge 1 \quad \text{hence} \quad \frac{v}{\sqrt{n}} \wedge 1 \leq B \leq \frac{(v + \sqrt{8E})}{\sqrt{n}} \wedge 1.$$

Using a change of variables, the definition of  $L$ , the monotonicity of  $\mathcal{H}$  and the bounds for  $B$ , we obtain that

$$\begin{aligned} E &\leq Cn^{1/2} \int_0^{2B} \sqrt{\mathcal{H}(1/z)} dz = Cn^{1/2} \int_{1/(2B)}^{+\infty} \frac{\sqrt{\mathcal{H}(u)}}{u^2} du \\ &\leq 2CLn^{1/2} B \sqrt{\mathcal{H}\left(\frac{1}{2B}\right)} \leq 2CL(v + \sqrt{8E}) \sqrt{H}. \end{aligned}$$

Solving this inequality with respect to  $E$  leads to the conclusion.

**9.5. Proof of Proposition 13.** Since

$$\mathbf{w}^S(\mathbf{s}, \bar{\mathbf{s}}, y) = \sqrt{2} \mathbb{E}_{\mathbf{s}} \left[ \sup_{t \in T} \left| \sum_{i=1}^n (U_{i,t} - \mathbb{E}[U_{i,t}]) \right| \right] \quad \text{with} \quad T = \mathcal{B}^S(\mathbf{s}, \bar{\mathbf{s}}, y)$$

and the  $U_{i,t}$  defined by (101), the result derives from the next proposition. In this case  $T \subset \mathcal{B}^S(\mathbf{s}, y)$  so that  $|T| \leq |\mathcal{B}^S(\mathbf{s}, y)|$ ,  $H = \log_+(2|\mathcal{B}^S(\mathbf{s}, y)|)$ ,  $b = \sqrt{2}$  and  $v^2 = 3y^2$  (because of Proposition 46).

**Proposition 50.** *Let  $T$  be a finite set and  $U_1, \dots, U_n$  independent random variables with values in  $\mathbb{R}^T$  satisfying for all  $t \in T$*

$$(111) \quad \max_{i=1, \dots, n} |U_{i,t}| \leq b \text{ a.s.}; \quad \sum_{i=1}^n \mathbb{E}[U_{i,t}^2] \leq v^2 \quad \text{and} \quad \log_+(2|T|) \leq H$$

for some positive numbers  $b, v$  and  $H$ . Then,

$$\mathbb{E} \left[ \sup_{t \in T} \left| \sum_{i=1}^n (U_{i,t} - \mathbb{E}[U_{i,t}]) \right| \right] \leq bH + v\sqrt{2H}.$$

*Proof.* Since the  $U_{i,t}$  are independent for  $i = 1, \dots, n$  and satisfy (111) for all  $t \in T$ , classical computations of the Laplace transform of  $S_{n,t} = \sum_{i=1}^n (U_{i,t} - \mathbb{E}[U_{i,t}])$  give, for  $\lambda \in (0, 1/b)$ ,

$$\mathbb{E} [\exp(\lambda |S_{n,t}|)] \leq 2 \exp \left[ \frac{\lambda^2 v^2}{2(1 - \lambda b)} \right] \quad \text{for all } t \in T.$$

For a proof of this inequality we refer to inequality (2.21) in Massart (2007). Applying Jensen's inequality and then this bound leads to

$$\begin{aligned} \mathbb{E} \left[ \sup_{t \in T} |S_{n,t}| \right] &= \frac{1}{\lambda} \log \left( \exp \left( \mathbb{E} \left[ \lambda \sup_{t \in T} |S_{n,t}| \right] \right) \right) \leq \frac{1}{\lambda} \log \mathbb{E} \left[ \exp \left( \lambda \sup_{t \in T} |S_{n,t}| \right) \right] \\ &\leq \frac{1}{\lambda} \log \left( \sum_{t \in T} \mathbb{E} [\exp(\lambda |S_{n,t}|)] \right) \leq \frac{H}{\lambda} + \frac{\lambda v^2}{2(1 - \lambda b)}. \end{aligned}$$

Minimizing the right-hand side with respect to  $\lambda \in (0, 1/b)$  leads to  $\lambda = (v + b\sqrt{2H})^{-1} \sqrt{2H}$  and finally  $\mathbb{E} [\sup_{t \in T} |S_{n,t}|] \leq bH + v\sqrt{2H}$ .  $\square$

**9.6. Proof of Proposition 17.** Let us denote by  $P_s$  the probability associated to  $s$  on  $(\mathcal{X}, \mathcal{A}, \mu)$ . For  $t = \sum_{I \in \mathcal{I}} [t_I / \mu(I)] \mathbb{1}_I$  with  $\mathbf{t} \in \mathcal{B}^S(\mathbf{s}, \bar{\mathbf{s}}, y)$ ,  $P_s$  almost surely,

$$\psi \left( \sqrt{\frac{t}{s}}(X_i) \right) = \sum_{I \in \mathcal{J}} \psi \left( \sqrt{\frac{t_I}{s_I}} \right) \mathbb{1}_I(X_i) \quad \text{for all } i = 1, \dots, n$$

and, by Cauchy-Schwarz Inequality,

$$\begin{aligned} S_n(t) &= \left| \sum_{i=1}^n \left[ \psi \left( \sqrt{\frac{t}{s}}(X_i) \right) - \mathbb{E}_{\mathbf{s}} \left[ \psi \left( \sqrt{\frac{t}{s}}(X_i) \right) \right] \right] \right| \\ &= \left| \sum_{I \in \mathcal{J}} \psi \left( \sqrt{\frac{t_I}{s_I}} \right) \sum_{i=1}^n [\mathbb{1}_I(X_i) - \mathbb{E}_{\mathbf{s}}[\mathbb{1}_I(X_i)]] \right| \\ &\leq \left[ \sum_{I \in \mathcal{J}} \psi^2 \left( \sqrt{\frac{t_I}{s_I}} \right) P_s(I) \right]^{1/2} \left[ \sum_{I \in \mathcal{J}} \left( \sum_{i=1}^n \frac{\mathbb{1}_I(X_i) - \mathbb{E}_{\mathbf{s}}[\mathbb{1}_I(X_i)]}{\sqrt{P_s(I)}} \right)^2 \right]^{1/2}. \end{aligned}$$

By Proposition 46, for all  $\mathbf{t} \in \mathcal{B}^S(\mathbf{s}, \bar{\mathbf{s}}, y)$ ,

$$n \mathbb{E}_{\mathbf{s}} \left[ \psi^2 \left( \sqrt{\frac{t}{s}}(X_1) \right) \right] = n \sum_{I \in \mathcal{I}} \psi^2 \left( \sqrt{\frac{t_I}{s_I}} \right) P_s(I) \leq 6y^2,$$

hence,  $P_s$  almost surely,

$$\sup_{\mathbf{t} \in \mathcal{B}^S(\mathbf{s}, \bar{\mathbf{s}}, y)} S_n(t) \leq \frac{y\sqrt{6}}{\sqrt{n}} \left[ \sum_{I \in \mathcal{J}} \left( \sum_{i=1}^n \frac{\mathbb{1}_I(X_i) - \mathbb{E}_{\mathbf{s}}[\mathbb{1}_I(X_i)]}{\sqrt{P_s(I)}} \right)^2 \right]^{1/2}.$$



Taking expectations on both sides and using the concavity of the square-root, we get

$$\mathbf{w}^S(\mathbf{s}, \bar{\mathbf{s}}, y) \leq \frac{y\sqrt{6}}{\sqrt{n}} \times \sqrt{\sum_{I \in \mathcal{I}} \sum_{i=1}^n \frac{P_s(I)}{P_s(I)}} = y\sqrt{6|\mathcal{I}|},$$

which leads to the result.

**9.7. Proof of Theorem 19.** In order to simplify the notations, when using the Hellinger distance on our model, we shall write  $h(\theta, \theta')$  instead of  $h(t_\theta, t_{\theta'})$ . All along this proof, we shall denote by  $|\cdot|$  the Euclidean distance on  $\mathbb{R}^d$  (as well as the absolute value when  $d = 1$ ) and by  $A_i$ ,  $2 \leq i \leq 9$ , constants that only depend on the structure of the parametric model  $\bar{S}$  as described by Assumption 18.

Since the parametric family  $\{t_\theta, \theta \in \Theta'\}$  is regular it has a continuous Fisher Information matrix  $I(\theta)$  which is also invertible on the compact set  $\Theta$  by Assumption 18-(ii). Therefore its eigenvalues are bounded away from zero and infinity on  $\Theta$  which implies — see (7.20) p.82 of the book by Ibragimov and Has'minskii (1981) — that

$$(112) \quad A_2 |\bar{\theta} - \theta| \leq h(\bar{\theta}, \theta) \leq A_3 |\bar{\theta} - \theta| \quad \text{with } 0 < A_2 < A_3 \quad \text{for all } \bar{\theta}, \theta \in \Theta.$$

It then follows from Assumption 18-(iii) that

$$(113) \quad \left\| \sqrt{\frac{t_\theta}{t_{\theta'}}} - \sqrt{\frac{t_{\bar{\theta}}}{t_{\bar{\theta}'}}} \right\|_\infty \leq \frac{A_1}{A_2} h(\theta, \bar{\theta}) = A_4 h(\theta, \bar{\theta}) \quad \text{for all } \theta, \bar{\theta} \text{ and } \theta' \in \Theta$$

and  $\bar{S}$  is therefore identifiable. Using the triangular inequality together with the facts that  $\psi$  is 1.15-Lipschitz and satisfies  $\psi(1/x) = -\psi(x)$  for all  $x > 0$ , we get for all  $\theta, \bar{\theta}, \theta', \bar{\theta}'$  in  $\Theta$ ,

$$(114) \quad \left\| \psi\left(\sqrt{\frac{t_{\theta'}}{t_\theta}}\right) - \psi\left(\sqrt{\frac{t_{\bar{\theta}'}}{t_{\bar{\theta}}}}\right) \right\|_\infty \leq 1.15 A_4 [h(\theta, \bar{\theta}) + h(\theta', \bar{\theta}')].$$

Moreover

**Lemma 51.** *The function*

$$(t, t') \mapsto \rho\left(t', \frac{t+t'}{2}\right) - \rho\left(t, \frac{t+t'}{2}\right)$$

*is uniformly continuous on  $\bar{S} \times \bar{S}$  with respect to the Hellinger distance.*

*Proof.* It is clearly enough to show the continuity of  $(t, t') \mapsto h(t, (t+t')/2)$  and, since

$$|h(t, (t+t')/2) - h(u, (u+u')/2)| \leq h(t, u) + h((t+t')/2, (u+u')/2),$$

it is enough to bound the second term. By the classical inequalities between the Hellinger and variation distances,

$$\begin{aligned} h^2\left(\frac{t+t'}{2}, \frac{u+u'}{2}\right) &\leq \frac{1}{2} \int \left| \frac{t+t'}{2} - \frac{u+u'}{2} \right| d\mu \\ &\leq \frac{1}{4} \int [|t-u| + |t'-u'|] d\mu \leq \frac{1}{\sqrt{2}} [h(t, u) + h(t', u')], \end{aligned}$$

which concludes the proof.  $\square$

Together with (112) and (114) the lemma shows that  $(\theta, \theta') \mapsto \mathbf{T}(\mathbf{X}, \mathbf{t}_\theta, \mathbf{t}_{\theta'})$  is continuous from  $\Theta \times \Theta$  into  $\mathbb{R}$  with probability 1, uniformly with respect to  $\mathbf{X}$ .

Recalling that  $s = t_\vartheta \in \overline{S}$ , let us set, for  $\Gamma \geq 1$ ,  $J \in \mathbb{N}$  and  $n \geq 1$ ,

$$\delta = \sqrt{\Gamma d/n} \quad \text{and} \quad \mathcal{C}(\Gamma, J) = \left\{ (\theta, \theta') \in \Theta^2 \text{ with } h(\vartheta, \theta) \leq \delta \text{ and } h(\vartheta, \theta') > 2^{J/2} \delta \right\}.$$

We want to establish the following intermediate result.

**Proposition 52.** *Under Assumption 18-(i),(ii) and (iii), there exist a positive constant  $C$  and a positive integer  $J_0$ , both depending on  $\overline{S}$  only, such that, for all  $J \geq J_0$  and  $\Gamma \in [1, n/d]$ ,*

$$\mathbb{P}_{\mathbf{s}} \left[ \sup_{(\theta, \theta') \in \mathcal{C}(\Gamma, J)} \mathbf{T}(\mathbf{X}, \mathbf{t}_\theta, \mathbf{t}_{\theta'}) < 0 \right] \geq 1 - \exp[-C2^J \Gamma d].$$

*Proof.* First of all, let us note that  $\sup_{(\theta, \theta') \in \mathcal{C}(\Gamma, J)} \mathbf{T}(\mathbf{X}, \mathbf{t}_\theta, \mathbf{t}_{\theta'})$  is measurable since  $\mathcal{C}(\Gamma, J)$  is separable and  $(\theta, \theta') \mapsto \mathbf{T}(\mathbf{X}, \mathbf{t}_\theta, \mathbf{t}_{\theta'})$  is continuous. Let us then set  $B = \{\theta \in \Theta \mid h(\vartheta, \theta) \leq \delta\}$ ,

$$C_j = \left\{ \theta \in \Theta \mid 2^{(J+j)/2} \delta \leq h(\vartheta, \theta) < 2^{(J+j+1)/2} \delta \right\} \quad \text{for all } j \in \mathbb{N}$$

and, for  $k \in \mathbb{N}$ , let  $B_k \subset B$  and  $C_{j,k} \subset C_j$  be  $2^{-k/2} \delta$ -nets for  $B$  and  $C_j$  respectively. Since, by (112),  $(\overline{S}, h)$  and  $(\Theta, |\cdot|) \subset (\mathbb{R}^d, |\cdot|)$  are isometric (up to constants), we may choose  $B_k$  and  $C_{j,k}$  in such a way that

$$(115) \quad \log |B_k| \leq A_5 dk \quad \text{and} \quad \log |C_{j,k}| \leq A_5 d(J + j + 1 + k) \quad \text{for all } k, j \in \mathbb{N},$$

as would be the case for Euclidean balls.

For  $\theta \in \Theta$  and  $j, k \in \mathbb{N}$ , we denote by  $\theta_k$  and  $\theta_{j,k}$  minimizers of the function  $\theta' \mapsto h(\theta, \theta')$  over  $B_k$  and  $C_{j,k}$  respectively. For all  $j \in \mathbb{N}$  and  $(\theta_0, \theta'_{j,0}) \in B_0 \times C_{j,0}$ , by Proposition 46,

$$\mathbb{E}_{\mathbf{s}} \left[ \psi^2 \left( \sqrt{t_{\theta'_{j,0}}/t_{\theta_0}} \right) \right] \leq 6 [h^2(\vartheta, \theta_0) + h^2(\vartheta, \theta'_{j,0})] \leq 6\delta^2 (1 + 2^{J+j+1}) \leq 2^{J+j+4} \delta^2.$$

Since  $\left| \psi \left( \sqrt{t_{\theta'_{j,0}}/t_{\theta_0}} \right) \right| \leq 1$ , we may use Bernstein's inequality with  $x_{j,0} = 2^{J+j} \Gamma d / 100$ , then (115) to derive that

$$\begin{aligned} & \mathbb{P}_{\mathbf{s}} \left[ \sup_{(\theta_0, \theta'_{j,0}) \in B_0 \times C_{j,0}} \mathbf{T}(\mathbf{X}, t_{\theta_0}, t_{\theta'_{j,0}}) - \mathbb{E}_{\mathbf{s}} [\mathbf{T}(\mathbf{X}, t_{\theta_0}, t_{\theta'_{j,0}})] > x_{j,0} \right] \\ & \leq \sum_{(\theta_0, \theta'_{j,0}) \in B_0 \times C_{j,0}} \mathbb{P}_{\mathbf{s}} \left[ \mathbf{T}(\mathbf{X}, t_{\theta_0}, t_{\theta'_{j,0}}) - \mathbb{E}_{\mathbf{s}} [\mathbf{T}(\mathbf{X}, t_{\theta_0}, t_{\theta'_{j,0}})] > x_{j,0} \right] \\ & = \sum_{(\theta_0, \theta'_{j,0}) \in B_0 \times C_{j,0}} \mathbb{P}_{\mathbf{s}} \left[ \sum_{i=1}^n \left( \psi \left( \sqrt{\frac{t_{\theta'_{j,0}}}{t_{\theta_0}}} (X_i) \right) - \mathbb{E}_{\mathbf{s}} \left[ \psi \left( \sqrt{\frac{t_{\theta'_{j,0}}}{t_{\theta_0}}} (X_i) \right) \right] \right) > \sqrt{2} x_{j,0} \right] \\ & \leq \exp \left[ A_5 d (J + j + 1) - \frac{2x_{j,0}^2}{2(2^{J+j+4} \Gamma d + \sqrt{2} x_{j,0}/3)} \right] \\ & \leq \exp [-C2^{J+j+1} \Gamma d] \leq \exp [-(j+1) - C2^J \Gamma d] \end{aligned}$$

for some  $C > 0$  and  $J_0$  large enough (depending on the  $A_i$ , which means on  $\overline{S}$ ) since  $J \geq J_0$ .

For  $(\theta, \theta') \in B \times C_j$  and  $k \in \mathbb{N}$ , let

$$\begin{aligned}
\Delta \mathbf{T} \left( \mathbf{X}, \mathbf{t}_{\theta_k}, \mathbf{t}_{\theta'_{j,k}}, \mathbf{t}_{\theta_{k+1}}, \mathbf{t}_{\theta'_{j,k+1}} \right) &= \left\{ \mathbf{T} \left( \mathbf{X}, \mathbf{t}_{\theta_{k+1}}, \mathbf{t}_{\theta'_{j,k+1}} \right) - \mathbb{E}_{\mathbf{s}} \left[ \mathbf{T} \left( \mathbf{X}, \mathbf{t}_{\theta_{k+1}}, \mathbf{t}_{\theta'_{j,k+1}} \right) \right] \right\} \\
&\quad - \left\{ \mathbf{T} \left( \mathbf{X}, \mathbf{t}_{\theta_k}, \mathbf{t}_{\theta'_{j,k}} \right) - \mathbb{E}_{\mathbf{s}} \left[ \mathbf{T} \left( \mathbf{X}, \mathbf{t}_{\theta_k}, \mathbf{t}_{\theta'_{j,k}} \right) \right] \right\} \\
&= \frac{1}{\sqrt{2}} \sum_{i=1}^n \left[ \psi \left( \sqrt{\frac{t_{\theta'_{j,k+1}}}{t_{\theta_{k+1}}}}(X_i) \right) - \psi \left( \sqrt{\frac{t_{\theta'_{j,k}}}{t_{\theta_k}}}(X_i) \right) \right] \\
&\quad - \frac{1}{\sqrt{2}} \sum_{i=1}^n \mathbb{E}_{\mathbf{s}} \left[ \psi \left( \sqrt{\frac{t_{\theta'_{j,k+1}}}{t_{\theta_{k+1}}}}(X_i) \right) - \psi \left( \sqrt{\frac{t_{\theta'_{j,k}}}{t_{\theta_k}}}(X_i) \right) \right].
\end{aligned}$$

It follows from (114) and (112) that

$$\begin{aligned}
\left\| \psi \left( \sqrt{\frac{t_{\theta'_{j,k+1}}}{t_{\theta_{k+1}}}} \right) - \psi \left( \sqrt{\frac{t_{\theta'_{j,k}}}{t_{\theta_k}}} \right) \right\|_{\infty} &\leq A_6 [h(\theta'_{j,k+1}, \theta'_{j,k}) + h(\theta_{k+1}, \theta_k)] \\
&\leq A_6 2^{1-k/2} \delta (2^{-1/2} + 1) < 7A_6 2^{-k/2-1} \delta,
\end{aligned}$$

therefore,

$$\mathbb{E}_{\mathbf{s}} \left[ \left[ \psi \left( \sqrt{\frac{t_{\theta'_{j,k+1}}}{t_{\theta_{k+1}}}} \right) - \psi \left( \sqrt{\frac{t_{\theta'_{j,k}}}{t_{\theta_k}}} \right) \right]^2 \right] < A_7 2^{-k} \delta^2.$$

For  $x_{j,k} = (k+1)2^{-k/2+J+j}\Gamma d/100$  and  $\Gamma d \leq n$ , we deduce from Bernstein's inequality and (115) that

$$\begin{aligned}
&\mathbb{P}_{\mathbf{s}} \left[ \sup_{\substack{(\theta_k, \theta_{k+1}) \in B_k \times B_{k+1} \\ (\theta'_{j,k}, \theta'_{j,k+1}) \in B_{j,k} \times B_{j,k+1}}} \Delta \mathbf{T} \left( \mathbf{X}, \mathbf{t}_{\theta_k}, \mathbf{t}_{\theta'_{j,k}}, \mathbf{t}_{\theta_{k+1}}, \mathbf{t}_{\theta'_{j,k+1}} \right) > x_{j,k} \right] \\
&\leq \exp \left[ 2A_5 d (J + j + 2k + 2) - \frac{2x_{j,k}^2}{2(A_7 2^{-k} \Gamma d + (7/6)x_{j,k} A_6 2^{-k/2} \delta)} \right] \\
&\leq \exp [-C((k+1)2^{J+j+1}\Gamma d)] \leq \exp [-(k+1) - (j+1) - C2^J \Gamma d],
\end{aligned}$$

for some  $C > 0$  and  $J_0$  large enough (depending on  $\overline{S}$ ).

Putting all these bounds together, we get, for  $J \geq J_0$  large enough and with probability at least

$$1 - e^{-C2^J \Gamma d} \left( \sum_{j \geq 1} e^{-j} + \sum_{j \geq 1} e^{-j} \sum_{k \geq 1} e^{-k} \right) \geq 1 - e^{-C2^J \Gamma d},$$

for some  $C > 0$ , that for all  $j \in \mathbb{N}$ ,  $\theta \in B$  and  $\theta' \in C_j$ ,

$$\begin{aligned}
\mathbf{T}(\mathbf{X}, \mathbf{t}_\theta, \mathbf{t}_{\theta'}) &= \mathbb{E}_{\mathbf{s}} [\mathbf{T}(\mathbf{X}, \mathbf{t}_\theta, \mathbf{t}_{\theta'})] + \mathbf{T}(\mathbf{X}, \mathbf{t}_\theta, \mathbf{t}_{\theta'}) - \mathbb{E}_{\mathbf{s}} [\mathbf{T}(\mathbf{X}, \mathbf{t}_\theta, \mathbf{t}_{\theta'})] \\
&= \mathbb{E}_{\mathbf{s}} [\mathbf{T}(\mathbf{X}, \mathbf{t}_\theta, \mathbf{t}_{\theta'})] + \lim_{k \rightarrow +\infty} \left\{ \mathbf{T}(\mathbf{X}, \mathbf{t}_{\theta_k}, \mathbf{t}_{\theta'_{j,k}}) - \mathbb{E}_{\mathbf{s}} [\mathbf{T}(\mathbf{X}, \mathbf{t}_{\theta_k}, \mathbf{t}_{\theta'_{j,k}})] \right\} \\
&= \mathbb{E}_{\mathbf{s}} [\mathbf{T}(\mathbf{X}, \mathbf{t}_\theta, \mathbf{t}_{\theta'})] + \mathbf{T}(\mathbf{X}, \mathbf{t}_{\theta_0}, \mathbf{t}_{\theta'_{j,0}}) - \mathbb{E}_{\mathbf{s}} [\mathbf{T}(\mathbf{X}, \mathbf{t}_{\theta_0}, \mathbf{t}_{\theta'_{j,0}})] \\
&\quad + \sum_{k \in \mathbb{N}} \Delta \mathbf{T}(\mathbf{X}, \mathbf{t}_{\theta_k}, \mathbf{t}_{\theta'_{j,k}}, \mathbf{t}_{\theta_{k+1}}, \mathbf{t}_{\theta'_{j,k+1}}) \\
&\leq \mathbb{E}_{\mathbf{s}} [\mathbf{T}(\mathbf{X}, \mathbf{t}_\theta, \mathbf{t}_{\theta'})] + x_{j,0} + \sum_{k \in \mathbb{N}} x_{j,k} \\
&\leq \mathbb{E}_{\mathbf{s}} [\mathbf{T}(\mathbf{X}, \mathbf{t}_\theta, \mathbf{t}_{\theta'})] + \frac{2^{J+j}\Gamma d}{100} \left( 1 + \sum_{k \geq 0} (k+1)2^{-k/2} \right).
\end{aligned}$$

Finally, with probability at least  $1 - e^{-C2^J\Gamma d}$ , for all  $(\theta, \theta') \in B \times C_j$  and  $j \in \mathbb{N}$ ,

$$(116) \quad \mathbf{T}(\mathbf{X}, \mathbf{t}_\theta, \mathbf{t}_{\theta'}) < \mathbb{E}_{\mathbf{s}} [\mathbf{T}(\mathbf{X}, \mathbf{t}_\theta, \mathbf{t}_{\theta'})] + 0.13 (2^{J+j}n\delta^2).$$

We conclude by using (13) which implies that, if  $(\theta, \theta') \in B \times C_j$ ,

$$\begin{aligned}
n^{-1}\mathbb{E}_{\mathbf{s}} [\mathbf{T}(\mathbf{X}, t_\theta, t_{\theta'})] &= \varrho(s, t_{\theta'}, t_\theta) - \varrho(s, t_\theta, t_{\theta'}) \leq \varrho(s, t_{\theta'}, t_\theta) - \rho(s, t_\theta) \\
&\leq \varrho(s, t_{\theta'}, t_\theta) - \rho(s, t_{\theta'}) + \rho(s, t_{\theta'}) - \rho(s, t_\theta) \\
&\leq \frac{1}{\sqrt{2}} [h^2(\vartheta, \theta') + h^2(\vartheta, \theta)] + h^2(\vartheta, \theta) - h^2(\vartheta, \theta') \\
&= - \left[ \left(1 - \frac{1}{\sqrt{2}}\right) h^2(\vartheta, \theta') - \left(1 + \frac{1}{\sqrt{2}}\right) h^2(\vartheta, \theta) \right] \\
&\leq - \left[ 2^{J+j} \left(1 - \frac{1}{\sqrt{2}}\right) - \left(1 + \frac{1}{\sqrt{2}}\right) \right] \delta^2 < -0.255 \times 2^{J+j} \delta^2
\end{aligned}$$

provided that  $J_0$  is large enough since  $J \geq J_0$ .  $\square$

Let us now proceed with the proof of Theorem 19. By Assumption 18-(iv), the MLE  $\tilde{\theta}_n$  converges towards the true parameter  $\vartheta$  and, since the model is regular, it converges at rate  $1/\sqrt{n}$  by Corollary 5.53 of van der Vaart (1998). Therefore, given  $\varepsilon > 0$ , for  $\Gamma$  large enough depending on  $\varepsilon$ ,  $h(\vartheta, \tilde{\theta}_n) \leq \delta = \sqrt{\Gamma d/n}$  with probability larger than  $1 - \varepsilon/2$ . We may now apply Proposition 52 with this particular value of  $\Gamma$ , provided that  $n \geq \Gamma d$ . It follows that, for a suitable choice of  $J \geq J_0$ ,

$$\mathbb{P}_{\mathbf{s}} \left[ \sup_{(\theta, \theta') \in \mathcal{C}(\Gamma, J)} \mathbf{T}(\mathbf{X}, \mathbf{t}_\theta, \mathbf{t}_{\theta'}) < 0 \right] \geq 1 - \exp[-C2^J\Gamma d] \geq 1 - \varepsilon/2.$$

Therefore

$$\mathbb{P}_{\mathbf{s}} \left[ \sup_{\theta' \in \mathcal{B}^c} \mathbf{T}(\mathbf{X}, \mathbf{t}_{\tilde{\theta}_n}, \mathbf{t}_{\theta'}) < 0 \right] \geq 1 - \varepsilon \quad \text{with} \quad \mathcal{B} = \left\{ \theta' \in \Theta \text{ such that } h(\vartheta, \theta') \leq 2^{J/2}\delta \right\}.$$

From now on, we shall work on the event of probability larger than  $1 - \varepsilon$  on which

$$(117) \quad h(\vartheta, \tilde{\theta}_n) \leq \delta \quad \text{and} \quad \sup_{\theta' \in \mathcal{B}^c} \mathbf{T}(\mathbf{X}, \mathbf{t}_{\tilde{\theta}_n}, \mathbf{t}_{\theta'}) < 0.$$

It remains to evaluate  $\sup_{\theta' \in \mathcal{B}} \mathbf{T}(\mathbf{X}, \mathbf{t}_{\tilde{\theta}_n}, \mathbf{t}_{\theta'})$  on this event. For all  $\theta' \in \mathcal{B}$ ,  $h(\tilde{\theta}_n, \theta') \leq 2^{(J+2)/2} \delta$ . Moreover, using the inequalities

$$0 \leq \sqrt{\frac{a+b}{2}} - \frac{\sqrt{a} + \sqrt{b}}{2} \leq \frac{(\sqrt{b} - \sqrt{a})^2}{4\sqrt{a}} \quad \text{for all } a, b > 0,$$

which both derive from  $2xy \leq x^2 + y^2$ , we get

$$\begin{aligned} & \rho\left(t_{\theta'}, \frac{t_{\tilde{\theta}_n} + t_{\theta'}}{2}\right) - \rho\left(t_{\tilde{\theta}_n}, \frac{t_{\tilde{\theta}_n} + t_{\theta'}}{2}\right) \\ &= \int (\sqrt{t_{\theta'}} - \sqrt{t_{\tilde{\theta}_n}}) \sqrt{\frac{t_{\tilde{\theta}_n} + t_{\theta'}}{2}} d\mu \\ &= \int (\sqrt{t_{\theta'}} - \sqrt{t_{\tilde{\theta}_n}}) \left[ \sqrt{\frac{t_{\tilde{\theta}_n} + t_{\theta'}}{2}} - \frac{\sqrt{t_{\theta'}} + \sqrt{t_{\tilde{\theta}_n}}}{2} \right] d\mu \\ &\leq \frac{1}{4} \int \left| \sqrt{\frac{t_{\theta'}}{t_{\theta'}}} - \sqrt{\frac{t_{\tilde{\theta}_n}}{t_{\theta'}}} \right| \times (\sqrt{t_{\theta'}} - \sqrt{t_{\tilde{\theta}_n}})^2 d\mu \\ &\leq \frac{1}{2} \left\| \sqrt{\frac{t_{\theta'}}{t_{\theta'}}} - \sqrt{\frac{t_{\tilde{\theta}_n}}{t_{\theta'}}} \right\|_{\infty} h^2(\tilde{\theta}_n, \theta') \leq \frac{A_4}{2} h^3(\tilde{\theta}_n, \theta') \end{aligned} \tag{118}$$

by (113). Besides, when  $u$  converges to 0,  $\psi(1+u) = (1/\sqrt{2}) \log(1+u) + O(u^3)$ . Setting  $u = \sqrt{t_{\theta'}/t_{\tilde{\theta}_n}} - 1$  so that by (113)  $|u| \leq A_4 h(\tilde{\theta}_n, \theta')$ , then using the fact that  $\tilde{\theta}_n$  maximizes the likelihood, we derive that

$$2\sqrt{2} \sum_{i=1}^n \psi\left(\sqrt{\frac{t_{\theta'}}{t_{\tilde{\theta}_n}}}(X_i)\right) \leq \sum_{i=1}^n \log t_{\theta'}(X_i) - \sum_{i=1}^n \log t_{\tilde{\theta}_n}(X_i) + A_8 n h^3(\tilde{\theta}_n, \theta') \leq A_8 n h^3(\tilde{\theta}_n, \theta').$$

Together with (117) and (118) this shows that, with probability larger than  $1 - \varepsilon$ ,

$$\sup_{\theta' \in \mathcal{B}} \mathbf{T}(\mathbf{X}, \mathbf{t}_{\tilde{\theta}_n}, \mathbf{t}_{\theta'}) \leq \frac{A_4 + A_8}{4} n h^3(\tilde{\theta}_n, \theta'),$$

hence

$$\sup_{\theta' \in \Theta} \mathbf{T}(\mathbf{X}, \mathbf{t}_{\tilde{\theta}_n}, \mathbf{t}_{\theta'}) < A_9 (2^J \Gamma d)^{3/2} n^{-1/2}. \tag{119}$$

Since the mapping  $(\theta, \theta') \mapsto \mathbf{T}(\mathbf{X}, t_{\theta}, t_{\theta'})$  is uniformly continuous on  $\Theta \times \Theta$ ,

$$\mathcal{E} = \left\{ t_{\theta} \in \overline{S}, \quad \sup_{\theta' \in \Theta} \mathbf{T}(\mathbf{X}, t_{\theta}, t_{\theta'}) < A_9 (2^J \Gamma d)^{3/2} n^{-1/2} \right\}$$

is an open subset of  $\overline{S}$  hence  $S \cap \mathcal{E}$  is also dense in  $\mathcal{E}$ . Besides,  $S \cap \mathcal{E} \subset \mathcal{E}(\mathbf{X}, S)$  for  $n$  large enough. Then, using (119) we get with probability at least  $1 - \varepsilon$

$$\mathbf{t}_{\tilde{\theta}_n} \in \mathcal{E} = \mathcal{E} \cap \text{Cl}(S \cap \mathcal{E}) \subset \text{Cl}(S \cap \mathcal{E}) \subset \text{Cl}(\mathcal{E}(\mathbf{X}, S)),$$

showing that  $\mathbf{t}_{\tilde{\theta}_n}$  is a  $\rho$ -estimator.

**9.8. Proof of Theorem 22.** Inequality (62) is obtained by combining (32) (Assumption 21-(i) corresponds to (31) in the density context) and (61). Consequently, it suffices to prove (61) and to do so we may assume with no loss of generality that  $s = \bar{s} \in S$ , which we shall do in the remaining part of this proof.

Let us consider the symmetric family  $\mathcal{F} = \mathcal{F}(y)$  defined for  $y = \sigma\sqrt{n} > 0$  by

$$\mathcal{F} = \mathcal{F}(y) = \left\{ \psi(\sqrt{t/s}) \mid \mathbf{t} \in \mathcal{B}^S(\mathbf{s}, y) \right\} \cup \left\{ -\psi(\sqrt{t/s}) \mid \mathbf{t} \in \mathcal{B}^S(\mathbf{s}, y) \right\}.$$

For all  $f \in \mathcal{F}$ ,  $|f| \leq 1$  and it follows from Proposition 46 that for all integers  $k \geq 2$ ,  $\mathbb{E}_{\mathbf{s}} [|f(X_1)|^k] \leq \mathbb{E}_{\mathbf{s}} [f^2(X_1)] \leq (6\sigma^2) \wedge 1$ . Since  $\psi$  is increasing and Lipschitz with Lipschitz constant  $L < \sqrt{3}$ , it follows from Assumption 21-(ii), (iii) that the family of pairs  $\mathcal{I}^\psi(s, \sigma, \epsilon)$  given by

$$\left\{ \left( \psi(\sqrt{t_L/s}), \psi(\sqrt{t_U/s}) \right), \left( -\psi(\sqrt{t_U/s}), -\psi(\sqrt{t_L/s}) \right), (t_L, t_U) \in \mathcal{I} \left( s, \sigma, L^{-1}\epsilon/\sqrt{2} \right) \right\}$$

covers  $\mathcal{F}$  with at most  $2 \exp \left[ \mathcal{H}_{[]}^S(s, \sigma, L^{-1}\epsilon/\sqrt{2}) \right] \leq \exp \left[ 2\mathcal{H}_{[]}^S(s, \sigma, L^{-1}\epsilon/\sqrt{2}) \right]$  brackets and that for all integers  $k \geq 2$

$$\begin{aligned} \mathbb{E}_{\mathbf{s}} \left[ \left( \psi(\sqrt{t_U/s})(X_1) - \psi(\sqrt{t_L/s})(X_1) \right)^k \right] \\ \leq 2^{k-2} \mathbb{E}_{\mathbf{s}} \left[ \left( \psi(\sqrt{t_U/s})(X_1) - \psi(\sqrt{t_L/s})(X_1) \right)^2 \right] \\ \leq 2^{k-2} L^2 \int_{\mathcal{X}} (\sqrt{t_U} - \sqrt{t_L})^2 d\mu \leq \epsilon^2 \times 2^{k-2} \leq \frac{k!}{2} \epsilon^2. \end{aligned}$$

Note that

$$\mathbf{w}^S(\mathbf{s}, \mathbf{s}, y) \leq \mathbb{E}_{\mathbf{s}} \left[ \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n (f(X_i) - \mathbb{E}_{\mathbf{s}} [f(X_i)]) \right| \right] = \mathbb{E}_{\mathbf{s}} \left[ \sup_{f \in \mathcal{F}} \left( \sum_{i=1}^n (f(X_i) - \mathbb{E}_{\mathbf{s}} [f(X_i)]) \right) \right].$$

We may therefore apply to this last expectation the bound (6.25) of Theorem 6.8 in Massart (2007) with  $\sigma^2$  replaced by  $(6\sigma^2) \wedge 1$ ,  $b = 1$ ,  $\delta = \epsilon$ ,  $H(\delta) = 2\mathcal{H}_{[]}^S(s, \sigma, L^{-1}\delta/\sqrt{2})$ ,  $\varepsilon = L/\sqrt{3} \in (0, 1]$  and  $A = \Omega$ . It leads to

$$\begin{aligned} \mathbf{w}^S(\mathbf{s}, \mathbf{s}, y) &\leq 27L^{-1}\sqrt{6n} \int_0^{\sigma L\sqrt{2}} \sqrt{\mathcal{H}_{[]}^S(s, \sigma, L^{-1}\epsilon/\sqrt{2})} d\epsilon + 8\mathcal{H}_{[]}^S(s, \sigma, \sigma\sqrt{3}/L) \\ &= 54\sqrt{3n} \int_0^\sigma \sqrt{\mathcal{H}_{[]}^S(s, \sigma, z)} dz + 8\mathcal{H}_{[]}^S(s, \sigma, \sigma\sqrt{3}/L). \end{aligned}$$

Since  $z \mapsto \mathcal{H}_{[]}^S(s, \sigma, z)$  is non-increasing,  $\mathcal{H}_{[]}^S(s, \sigma, \sigma\sqrt{3}/L) \leq \mathcal{H}_{[]}^S(s, \sigma, \sigma) \leq \sigma^{-2}\phi^2(\sigma)$ . Let us now choose some  $\lambda_0 > 1$ . It follows from the definition of  $\tau_n$  and the monotonicity of  $\sigma \mapsto \phi(\sigma)/\sigma$  that for all  $\lambda' \in ]1, \lambda_0]$  and  $\sigma \geq \lambda_0\tau_n$

$$\frac{\phi(\sigma)}{\sigma} \leq \frac{\phi(\lambda'\tau_n)}{\lambda'\tau_n} \leq \lambda'\tau_n\sqrt{n} \leq \frac{\lambda'}{\lambda_0}\sigma\sqrt{n}.$$

Letting  $\lambda'$  tend to 1 we get  $\phi(\sigma) \leq \sigma^2\sqrt{n}/\lambda_0$ . Putting these bounds together we get that, for all  $y = \sqrt{n}\sigma \geq \lambda_0\sqrt{n}\tau_n$  with  $\lambda_0 = 2555$ ,

$$\mathbf{w}^S(\mathbf{s}, \mathbf{s}, y) \leq 54\sqrt{3n}\phi(\sigma) + 8\sigma^{-2}\phi^2(\sigma) \leq \left( \frac{54\sqrt{3}}{\lambda_0} + \frac{8}{\lambda_0^2} \right) n\sigma^2 \leq c_0 y^2.$$

Finally,  $\sup_{\mathbf{s} \in S} D^S(\mathbf{s}, \mathbf{s}) \leq (\lambda_0^2 n \tau_n^2) \vee 1$ .

**9.9. Proof of Proposition 30.** If  $s \neq \bar{s}$ , let  $J \in \mathbb{N}$  be such that  $h^2(s, \bar{s}) = 2^{-J}$  and  $\Omega_J(\mathbf{X}) = \{\omega \in \Omega \mid \ell(X_i) \leq J \text{ for } i = 1, \dots, n\}$ . Since  $s \in \bar{S}$ , there exists  $\theta^* \in \Theta$  such that  $s = s_{\theta^*}$  and for  $y \geq 1$ , let us set

$$\Theta[\theta^*, y] = \{\theta \in \Theta \mid \mathbf{s}_\theta \in \mathcal{B}^S(\mathbf{s}, \bar{s}, y)\}.$$

We decompose  $\mathbf{w}^S(\mathbf{s}, \bar{s}, y)$  as

$$\mathbf{w}^S(\mathbf{s}, \bar{s}, y) = \mathbb{E}_{\mathbf{s}} \left[ \sup_{\mathbf{t} \in \mathcal{B}^S(\mathbf{s}, \bar{s}, y)} |\mathbf{Z}(\mathbf{X}, \bar{s}, \mathbf{t})| \right] = \mathbb{E}_{\mathbf{s}} \left[ \sup_{\theta \in \Theta[\theta^*, y]} |\mathbf{Z}(\mathbf{X}, \bar{s}, \mathbf{s}_\theta)| \right] = E_1 + E_2$$

with

$$E_1 = \mathbb{E}_{\mathbf{s}} \left[ \sup_{\theta \in \Theta[\theta^*, y]} |\mathbf{Z}(\mathbf{X}, \bar{s}, \mathbf{s}_\theta)| \mathbb{1}_{\Omega_J(\mathbf{X})} \right] \quad \text{and} \quad E_2 = \mathbb{E}_{\mathbf{s}} \left[ \sup_{\theta \in \Theta[\theta^*, y]} |\mathbf{Z}(\mathbf{X}, \bar{s}, \mathbf{s}_\theta)| \mathbb{1}_{(\Omega_J(\mathbf{X}))^c} \right].$$

Let us bound each of these last two terms from above. On the event  $\Omega_J(\mathbf{X})$ ,  $s(X_i) = \bar{s}(X_i)$  for all  $i = 1, \dots, n$ , hence

$$E_1 \leq \mathbb{E}_{\mathbf{s}} \left[ \sup_{\theta \in \Theta[\theta^*, y]} \left| \sum_{i=1}^n \left( \psi \left( \sqrt{\frac{s_\theta(X_i)}{s(X_i)}} \right) - \mathbb{E}_{\mathbf{s}} \left[ \psi \left( \sqrt{\frac{s_\theta(X_i)}{s(X_i)}} \right) \right] \right) \right| \right].$$

For  $\theta \in \Theta$  and  $i = 1, \dots, n$ , either  $\ell(X_i) \leq J(\theta, \theta^*)$  and  $s_\theta(X_i) = s(X_i) = 2^{-\ell(X_i)}$  in which case,  $\psi \left( \sqrt{s_\theta(X_i)/s(X_i)} \right) = \psi(1) = 0$ , or  $\ell(X_i) > J(\theta, \theta^*)$ ,  $s_\theta(X_i) = 0$  and then  $\psi \left( \sqrt{s_\theta(X_i)/s(X_i)} \right) = \psi(0) = -1$ . In both cases  $\psi \left( \sqrt{s_\theta(X_i)/s(X_i)} \right) = -\mathbb{1}_{\ell(X_i) > J(\theta, \theta^*)}$ . Let us now introduce  $n$  Rademacher random variables  $\varepsilon_1, \dots, \varepsilon_n$ , independent of the  $X_i$ . By a symmetrization argument,

$$E_1 \leq 2\mathbb{E}_{\mathbf{s}} \left[ \sup_{\theta \in \Theta[\theta^*, y]} \left| \sum_{i=1}^n \varepsilon_i \psi \left( \sqrt{\frac{s_\theta(X_i)}{s(X_i)}} \right) \right| \right] = 2\mathbb{E}_{\mathbf{s}} \left[ \sup_{\theta \in \Theta[\theta^*, y]} \left| \sum_{i=1}^n \varepsilon_i \mathbb{1}_{\ell(X_i) > J(\theta, \theta^*)} \right| \right].$$

Let us now work conditionally on  $X_1, \dots, X_n$  and denote by  $\mathbb{E}_\varepsilon$  the corresponding conditional expectation. Up to a re-ordering of the  $\varepsilon_i$ , we may assume with no loss of generality that  $\ell(X_1) \geq \ell(X_2) \geq \dots \geq \ell(X_n)$ . Then  $\sum_{i=1}^n \varepsilon_i \mathbb{1}_{\ell(X_i) > J(\theta, \theta^*)}$  is necessarily of the form  $\sum_{i=1}^k \varepsilon_i$  for some non-negative integer  $k = k(\theta, \theta^*, \mathbf{X})$  corresponding to the number of  $X_i$  of length  $\ell(X_i)$  larger than  $J(\theta, \theta^*)$ . By (72), for  $\theta \in \Theta[\theta^*, y]$ ,  $J(\theta, \theta^*) \geq \log_2(n/y^2)$ , hence  $k$  cannot exceed the number  $\hat{N}$  of  $X_i$  of length not smaller than  $\log_2(n/y^2)$ . We deduce that

$$\sup_{\theta \in \Theta[\theta^*, y]} \left| \sum_{i=1}^n \varepsilon_i \mathbb{1}_{\ell(X_i) > J(\theta, \theta^*)} \right| \leq \max_{0 \leq k \leq \hat{N}} \left| \sum_{i=1}^k \varepsilon_i \right|$$



with the convention  $\sum_{i=1}^0 = 0$ . Taking the expectation (conditionnaly on  $\mathbf{X}$ ) and using Doob's maximal inequality we get

$$\begin{aligned} \mathbb{E}_\varepsilon \left[ \sup_{\theta \in \Theta[\theta^*, y]} \left| \sum_{i=1}^n \varepsilon_i \mathbb{1}_{\ell(X_i) > J(\theta, \theta^*)} \right| \right] \\ \leq \mathbb{E}_\varepsilon \left[ \max_{0 \leq k \leq \widehat{N}} \left| \sum_{i=1}^k \varepsilon_i \right| \right] \leq \left( \mathbb{E}_\varepsilon \left[ \max_{0 \leq k \leq \widehat{N}} \left| \sum_{i=1}^k \varepsilon_i \right|^2 \right] \right)^{1/2} \\ \leq 2 \left( \mathbb{E}_\varepsilon \left[ \left| \sum_{i=1}^{\widehat{N}} \varepsilon_i \right|^2 \right] \right)^{1/2} = 2\sqrt{\widehat{N}}. \end{aligned}$$

Taking the expectation with respect to  $X_1, \dots, X_n$  finally leads to

$$\begin{aligned} E_1 &\leq 4\mathbb{E}_\mathbf{s} \left[ \sqrt{\widehat{N}} \right] \leq 4\sqrt{\mathbb{E}_\mathbf{s} \left( \sum_{i=1}^n \mathbb{1}_{\ell(X_i) > \log_2(n/y^2)} \right)} \\ &= 4\sqrt{n\mathbb{P}_\mathbf{s} \left[ \ell(X_1) > \log_2 \left( \frac{n}{y^2} \right) \right]} = 4 \left[ n \sum_{j > \log_2(n/y^2)} 2^{-j} \right]^{1/2} \leq 4\sqrt{2y^2}. \end{aligned}$$

As to  $E_2$ , since  $\psi$  is bounded by one and  $y \geq 1$ , it satisfies

$$E_2 \leq 2n\mathbb{P}_\mathbf{s} [\Omega_J(\mathbf{X})^c] \leq 2n^2\mathbb{P}_\mathbf{s} [\ell(X_1) > J] = 2n^2 2^{-J} \leq 2n^2 2^{-J} y = 2n^2 h^2(s, \bar{s}) y.$$

Putting these bounds together, we get

$$(120) \quad \mathbf{w}^S(\mathbf{s}, \bar{\mathbf{s}}, y) = E_1 + E_2 \leq 2 \left( 2\sqrt{2} + n^2 h^2(s, \bar{s}) \right) y \quad \text{for all } y \geq 1,$$

which leads to the bound on  $D^S(\mathbf{s}, \bar{\mathbf{s}})$ .

If  $s = \bar{s}$ , we proceed in the same way with  $J = +\infty$  which means that  $\Omega_J(\mathbf{X})^c = \emptyset$ ,  $E_2 = 0$  and (120) remains valid.

**9.10. Proof of Proposition 37.** Note that the set of densities  $\{p^\beta, \beta > 0\}$  is a regular statistical model with respect to the parameter  $\beta$ . Following Theorem 2.1 p. 121 (equation 2.9) and Section 5 p.133 of the book by Ibragimov and Has'minskiĭ (1981), for  $\beta' > \beta > 0$

$$(121) \quad h^2(p^\beta, p^{\beta'}) \leq \left[ \sup_{\beta \leq b \leq \beta'} \frac{I(b)(\beta' - \beta)^2}{8} \right] \wedge 1,$$

where  $I$  denotes the Fisher Information of this parametric model which is given by

$$I(b) = \int_{\mathbb{R}} \frac{[\dot{p}^b(x)]^2}{p^b(x)} dx = 2 \int_0^\infty \frac{[\dot{p}^b(x)]^2}{p^b(x)} dx$$

and  $\dot{p}^b(x)$  is the derivative of  $p^b(x)$  with respect to  $b$ . It follows from (88) that

$$\dot{p}^b(x) = p^b(x) \left[ -\frac{1}{b} - \frac{\Gamma'(b)}{\Gamma(b)} + \frac{1}{b^2} x^{1/b} \log x \right] \quad \text{for } x > 0,$$

hence

$$\frac{[p^b(x)]^2}{p^b(x)} \leq 3p^b(x) \left[ \frac{1}{b^2} + \left( \frac{\Gamma'(b)}{\Gamma(b)} \right)^2 + \frac{1}{b^4} x^{2/b} (\log x)^2 \right]$$

and

$$I(b) \leq 3 \left[ \frac{1}{b^2} + \left( \frac{\Gamma'(b)}{\Gamma(b)} \right)^2 + \frac{1}{b^5 \Gamma(b)} \int_0^\infty e^{-x^{1/b}} x^{2/b} (\log x)^2 dx \right].$$

Using again a change of variables we get

$$\int_0^\infty e^{-x^{1/b}} x^{2/b} (\log x)^2 dx = b^3 \int_0^\infty e^{-u} u^{b+1} (\log u)^2 du = b^3 \Gamma''(b+2),$$

so that finally,

$$I(b) \leq 3J \quad \text{with} \quad J \leq \frac{1}{b^2} + \left( \frac{\Gamma'(b)}{\Gamma(b)} \right)^2 + \frac{\Gamma''(b+2)}{b^2 \Gamma(b)} = \frac{1}{b^2} + \left( \frac{\Gamma'(b)}{\Gamma(b)} \right)^2 + \frac{(b+1)\Gamma''(b+2)}{b\Gamma(b+2)}.$$

Binet's formula for  $\log \Gamma$  (see Whittaker and Watson (1996) page 251) tells us that

$$(122) \quad \frac{\Gamma'(b)}{\Gamma(b)} = \log b - \frac{1}{2b} - 2k(b) \quad \text{with} \quad k(b) = \int_0^\infty \frac{x}{(x^2 + b^2)(e^{2\pi x} - 1)} dx$$

hence

$$\frac{\Gamma''(b)}{\Gamma(b)} - \left( \frac{\Gamma'(b)}{\Gamma(b)} \right)^2 = \frac{1}{b} + \frac{1}{2b^2} + 4b \int_0^\infty \frac{x}{(x^2 + b^2)^2 (e^{2\pi x} - 1)} dx.$$

One should then observe that, since  $e^u \geq 1 + u$ ,

$$(123) \quad 0 \leq 2k(b) \leq \frac{1}{\pi} \int_0^\infty \frac{dx}{(x^2 + b^2)} = \frac{1}{2b}$$

and

$$0 \leq \int_0^\infty \frac{x}{(x^2 + b^2)^2 (e^{2\pi x} - 1)} dx \leq \frac{1}{2\pi} \int_0^\infty \frac{dx}{(x^2 + b^2)^2} = \frac{1}{2\pi b^3} \int_0^\infty \frac{dx}{(x^2 + 1)^2} = \frac{1}{8b^3}.$$

It follows that

$$(124) \quad J \leq \frac{1}{b^2} + \left( \frac{\Gamma'(b)}{\Gamma(b)} \right)^2 + \frac{b+1}{b} \left[ \left( \frac{\Gamma'(b+2)}{\Gamma(b+2)} \right)^2 + \frac{1}{b+2} + \frac{1}{(b+2)^2} \right].$$

Moreover, since  $\log \Gamma$  is a strictly convex function (Whittaker and Watson (1996) page 250) with a minimum value at  $b_0 \in (1, 2)$ , by (122) and (123),  $0 < \Gamma'(b)/\Gamma(b) < \log b - (2b)^{-1}$  for  $b > b_0$  and  $0 < -\Gamma'(b)/\Gamma(b) < b^{-1} - \log b$  for  $b < b_0$ . It follows that

$$\left| \frac{\Gamma'(b)}{\Gamma(b)} \right| \leq \begin{cases} 1.37b^{-1} & \text{if } 0 < b \leq 1; \\ 1 & \text{if } 1 < b \leq 3; \\ \log b - (2b)^{-1} & \text{if } b > 3. \end{cases}$$

Therefore, by (124),

$$J \leq \frac{2.88}{b^2} + \frac{b(b+1)}{b^2} \left[ 1 + \frac{1}{b+2} + \frac{1}{(b+2)^2} \right] \leq \frac{52}{9b^2} \quad \text{for } b \leq 1;$$

$$J \leq \frac{1}{b^2} + 1 + \frac{b+1}{b} \left[ \left( \log(b+2) - \frac{1}{2(b+2)} \right)^2 + \frac{1}{b+2} + \frac{1}{(b+2)^2} \right] \leq 4.63 \quad \text{for } 1 < b \leq 3$$

and

$$J \leq \frac{1}{b^2} + (\log b)^2 + \frac{b+1}{b} \left[ [\log(b+2)]^2 + \frac{1}{b+2} + \frac{1}{(b+2)^2} \right] \leq 4.23(\log b)^2 \quad \text{for } b > 3.$$

Finally

$$\frac{I(b)}{8} \leq \begin{cases} 13/(6b^2) & \text{if } 0 < b \leq 1 \\ 7/4 & \text{if } 1 < b \leq 3 \\ (1.3 \log b)^2 & \text{if } b > 3 \end{cases}$$

and our first bound then follows from (121).

Let us now turn to the second inequality.

$$h^2(p^\beta, p^0) = 1 - \int_{-1}^1 \sqrt{p^\beta(x)/2} dx = 1 - \frac{1}{\sqrt{\beta\Gamma(\beta)}} \int_0^1 \exp\left[-\frac{x^{1/\beta}}{2}\right] dx.$$

Since  $\beta\Gamma(\beta) = \Gamma(\beta+1) \leq 1$  for  $0 < \beta \leq 1$  and

$$\int_0^1 \exp\left[-\frac{x^{1/\beta}}{2}\right] dx \geq 1 - \int_0^1 \frac{x^{1/\beta}}{2} dx = 1 - \frac{\beta}{2(\beta+1)},$$

(90) follows.

To control  $\mathbf{w}_{p^\beta}$  we observe that, for  $\beta > 0$ ,

$$h^2\left(p^\beta, \frac{1}{\lambda} p^\beta\left(\frac{\cdot}{\lambda}\right)\right) = 1 - \frac{1}{\beta\Gamma(\beta)\sqrt{\lambda}} \int_0^{+\infty} e^{-(1/2)x^{1/\beta}(1+\lambda^{-1/\beta})} dx.$$

Using the change of variables  $z = x((1+\lambda^{-1/\beta})/2)^\beta$ , and the assumption  $\lambda \in [1, 2]$ , we get

$$\begin{aligned} h^2\left(p^\beta, \frac{1}{\lambda} p^\beta\left(\frac{\cdot}{\lambda}\right)\right) &= 1 - \frac{1}{\beta\Gamma(\beta)\sqrt{\lambda}} \int_0^{+\infty} 2^\beta (1+\lambda^{-1/\beta})^{-\beta} e^{-z^{1/\beta}} dz \\ &= 1 - \frac{2^\beta}{\sqrt{\lambda} (1+\lambda^{-1/\beta})^\beta} = \left(\frac{2}{\lambda^{1/\beta}+1}\right)^\beta \left[\left(\frac{\lambda^{1/\beta}+1}{2}\right)^\beta - \sqrt{\lambda}\right] \\ &\leq \left(\frac{\lambda^{1/\beta}+1}{2}\right)^\beta - \sqrt{\lambda} \leq \lambda - \sqrt{\lambda} < (3/5)(\lambda-1). \end{aligned}$$

The particular case of  $\beta = 0$  is straightforward.

**Acknowledgements.** One of the authors is grateful to Vladimir Koltchinskii for stimulating discussions and especially letting him know about the nice properties of VC-subgraph classes and all authors would like to thank the referee for his/her many useful comments.

## REFERENCES

- Audibert, J.-Y. and Catoni, O. (2011). Robust linear least squares regression. *Ann. Statist.*, 39(5):2766–2794.
- Baraud, Y. (2002). Model selection for regression on a random design. *ESAIM Probab. Statist.*, 6:127–146.
- Baraud, Y. (2011). Estimator selection with respect to Hellinger-type risks. *Probab. Theory Related Fields*, 151(1-2):353–401.

- Barron, A., Birgé, L., and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413.
- Barron, A. R. (1991). Complexity regularization with application to artificial neural networks. In *Nonparametric Functional Estimation and Related Topics (Spetses, 1990)*, volume 335 of *NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci.*, pages 561–576. Kluwer Acad. Publ., Dordrecht.
- Birgé, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete*, 65(2):181–237.
- Birgé, L. (1984). Stabilité et instabilité du risque minimax pour des variables indépendantes équidistribuées. *Ann. Inst. H. Poincaré Probab. Statist.*, 20(3):201–223.
- Birgé, L. (2006). Model selection via testing: an alternative to (penalized) maximum likelihood estimators. *Ann. Inst. H. Poincaré Probab. Statist.*, 42(3):273–325.
- Birgé, L. (2013). Robust tests for model selection. In Banerjee, M., Bunea, F., Huang, J., Koltchinskii, V., and Maathuis, M. H., editors, *From Probability to Statistics and Back: High-Dimensional Models and Processes*, volume 9, pages 47–64. IMS Collections.
- Birgé, L. and Massart, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields*, 97(1-2):113–150.
- Birgé, L. and Massart, P. (1997). From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*, pages 55–87. Springer, New York.
- Birgé, L. and Massart, P. (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375.
- Birgé, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2):33–73.
- Dudley, R. M. (1984). A course on empirical processes. In *École d'été de Probabilités de Saint-Flour, XII—1982*, volume 1097 of *Lecture Notes in Math.*, pages 1–142. Springer, Berlin.
- Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531.
- Giné, E. and Koltchinskii, V. (2006). Concentration inequalities and asymptotic results for ratio type empirical processes. *Ann. Probab.*, 34(3):1143–1216.
- Grenander, U. (1981). *Abstract inference*. John Wiley & Sons, Inc., New York. Wiley Series in Probability and Mathematical Statistics.
- Hájek, J. (1972). Local asymptotic minimax and admissibility in estimation. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971)*, Vol. I: *Theory of statistics*, pages 175–194. Univ. California Press, Berkeley, Calif.
- Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.*, 35:73–101.
- Huber, P. J. (1981). *Robust Statistics*. John Wiley & Sons, Inc., New York. Wiley Series in Probability and Mathematical Statistics.
- Ibragimov, I. A. and Has'minskiĭ, R. Z. (1980). On estimate of the density function. *Zap. Nauchn. Semin. LOMI*, 98(61–85).
- Ibragimov, I. A. and Has'minskiĭ, R. Z. (1981). *Statistical Estimation. Asymptotic Theory*, volume 16. Springer-Verlag, New York.
- Klein, T. and Rio, E. (2005). Concentration around the mean for maxima of empirical processes. *Ann. Probab.*, 33(3):1060–1077.

- Kolmogorov, A. N. and Tihomirov, V. M. (1961).  $\varepsilon$ -entropy and  $\varepsilon$ -capacity of sets in functional space. *Amer. Math. Soc. Transl. (2)*, 17:277–364.
- Koltchinskii, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.*, 34(6):2593–2656.
- Le Cam, L. (1970). On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *Ann. Math. Statist.*, 41:802–828.
- Le Cam, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.*, 1:38–53.
- Le Cam, L. (1975). On local and global properties in the theory of asymptotic normality of experiments. In *Stochastic processes and related topics (Proc. Summer Res. Inst. Statist. Inference for Stochastic Processes, Indiana Univ., Bloomington, Ind., 1974, Vol. 1; dedicated to Jerzy Neyman)*, pages 13–54. Academic Press, New York.
- Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer Series in Statistics. Springer-Verlag, New York.
- Le Cam, L. (1990). Maximum likelihood: An introduction. *Inter. Statist. Review*, 58(2):153–171.
- Le Cam, L. and Yang, G. L. (1990). *Asymptotics in Statistics. Some Basic Concepts*. Springer Series in Statistics. Springer-Verlag, New York.
- Massart, P. (2007). *Concentration Inequalities and Model Selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003.
- Massart, P. and Nédélec, É. (2006). Risk bounds for statistical learning. *Ann. Statist.*, 34(5):2326–2366.
- Sart, M. (2014). Estimation of the transition density of a markov chain. *Annales de l’I.H.P. Probabilités et statistiques*, 50(3):1028–1068.
- Sart, M. (2015). Model selection for poisson processes with covariates. *ESAIM: PS*, 19:204–235.
- Sart, M. (2016). Robust estimation on a parametric model via testing. *Bernoulli*, 22(3):1617–1670.
- van de Geer, S. (1995). The method of sieves and minimum contrast estimators. *Math. Methods Statist.*, 4(1):20–38.
- van der Vaart, A. and Wellner, J. A. (2009). A note on bounds for VC dimensions. In *High Dimensional Probability V: the Luminy volume*, volume 5 of *Inst. Math. Stat. Collect.*, pages 103–107. Inst. Math. Statist., Beachwood, OH.
- van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.
- van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes. With Applications to Statistics*. Springer Series in Statistics. Springer-Verlag, New York.
- Whittaker, E. T. and Watson, G. N. (1996). *A Course of Modern Analysis*. Cambridge Mathematical Library. Cambridge University Press, Cambridge. An introduction to the general theory of infinite processes and of analytic functions; with an account of the principal transcendental functions, Reprint of the fourth (1927) edition.
- Yang, Y. and Barron, A. (1999). Information-theoretic determination of minimax rates of convergence. *Ann. Statist.*, 27(5):1564–1599.

UNIV. NICE SOPHIA ANTIPOLIS, CNRS, LJAD, UMR 7351, 06100 NICE, FRANCE.

*E-mail address:* `baraud@unice.fr`

SORBONNE UNIVERSITÉS, UPMC UNIV. PARIS 06, CNRS - UMR 7599, LPMA - CASE COURRIER 188,  
75252 PARIS CEDEX 05, FRANCE.

*E-mail address:* `lucien.birge@upmc.fr`

UNIV LYON, UJM-SAINT-ETIENNE, CNRS, INSTITUT CAMILLE JORDAN UMR 5208, F-42023, SAINT-  
ETIENNE, FRANCE.

*E-mail address:* `mathieu.sart@univ-st-etienne.fr`